Identification of Groupings of Graph Theoretical Molecular Descriptors Using a Hybrid Cluster Analysis Approach

Stavros L. Taraviras,[†,‡] Ovidiu Ivanciuc,*[,§,||] and Daniel Cabrol-Bass*[,†,⊥]

Chimiometry and Molecular Modeling Group, Arômes, Synthèses et Interactions Lab, University of Nice-Sophia Antipolis, 28 Avenue Valrose, 06108 Nice Cedex 2, France, and Department of Organic Chemistry, Faculty of Chemical Technology, University "Politehnica" of Bucharest, Oficiul 12 CP 243, 78100 Bucharest, Romania

Received December 16, 1999

There is an abundance of structural molecular descriptors of various forms that have been proposed and tested over the years. Very often different descriptors represent, more or less, the same aspects of molecular structures and, thus, they have diminished discriminating power for the identification of different structural features that might contribute to the molecular property, or activity of interest. Therefore, it is essential that noncorrelated descriptors be employed to ensure the wider and the less inflated possible coverage of the chemical space. The most usual approach for reducing the number of descriptors and employing noncorrelated (or orthogonal) descriptors involves principal component analysis (PCA) or other factor analytical techniques. In this work we present an approach for determining relationships (groupings) among 240 graph-theoretical descriptors, as a means for selecting nonredundant ones, based on the application of cluster analysis (CA). To remove inherent biases and particularities of different CA algorithms, several clustering solutions, using these algorithms, were "hybridized" to obtain a reliable and confident overall solution concerning how the interrelationships within the data are structured. The calculated correlation coefficients between descriptors were used as a reference for a discussion on the different CA methods employed, and the resulted clusters of descriptors were statistically analyzed for deriving the intercorrelations between the different operators, weighting schemes and matrices used for the computation of these descriptors.

INTRODUCTION

Molecular Diversity and Molecular Descriptors. Combinatorial chemistry was introduced, as an alternative to the rational molecular design approach, to allow the synthesis of a wide range of specific and diverse compounds. These compounds are then screened quickly and reliably against a plethora of biological, or other, targets as part of the molecular discovery process. Under this rationale, it would seem to be the ideal way to identify new lead structures if the molecules synthesized and the ones chosen for screening represent an evenly distributed cross-section of all potentially bioactive (for the case of pharmaceutical compounds) structures. This requirement is the basis for the concept of molecular diversity. The obvious approach is to characterize the chemical structures in terms of molecular descriptors based on their physicochemical or biological properties, their substructures, or any other possibly relevant and calculable features. Then one needs only to require that the ranges and occurrences of these various features adequately describe the universe of molecules, that is, cover the chemical space of interest. This is the foundation of the similar property principle, which states that chemically (structurally) similar molecules will be expected to (often) exhibit similar biological and physicochemical profiles.[1] The standard procedure followed involves the computation of molecular descriptors selected from classes of quantifiable two-dimensional (2D) and three-dimensional (3D) molecular parameters. Examples of the former include the atom pairs, molecular fingerprints, and topological indices, whereas examples of the latter include receptor recognition features, molecular shape, and pharmacophores.[2−10] Generally, the 2D structural descriptors are considered to be incorporating the major part of the essential information about the chemical structure, and they have been found to perform better than or at least as equally well as the most complicated and resource-demanding 3D descriptors, in most of the comparative studies conducted thus far.[3−16] In addition the 2D molecular structure is the essential basis of chemical synthesis strategies, both chemically and conceptually, and the same applies to patent claims, where the molecules are treated as 2D objects. For assessing the diversity of a collection of pharmacologically important compounds, we must bear in mind that these compounds have certain characteristics which facilitate their participation in 3D physiological/biochemical events, namely, they are small in size and molecular weight and they do not contain "exotic" compounds, properties which go along well with the 2D molecular description.

Topological Indices. The topological indices (TIs) are 2D molecular descriptors whose values are associated with the structural constitution of a chemical compound. They are computed directly from the molecular graph of the particular compound, and they offer several important advantages: they

* Corresponding author.
[†] University of Nice-Sophia Antipolis.
[‡] Present address: SGI (Silicon Graphics), European HQ, 1530 Arlington Business Park, Reading, RG7 4SB, U.K.
[§] University "Politehnica" of Bucharest.
[||] E-mail: o_ivancicu@chim.upb.ro.
[⊥] Telephone: +33 (0)4 92 07 61 20. E-mail: cabrol@unice.fr.

10.1021/ci990149y CCC: $19.00
Published on Web 08/26/2000

are calculated with minimal computational resources, they do not require geometrical optimization of the 3D molecular structure which is computationally demanding, complex, and often unreliable, and they have a unique value for a particular chemical compound providing a description of the entire molecule. Being real numerical values, TIs can be easily visualized, manipulated, and statistically evaluated, and missing values can be replaced. Also, they can capture the structural information of even odd compounds, whereas, say, fingerprints may fail to encode rare chemical functionalities, since they depend on predefined dictionaries generated from already known molecules. Last, when applied to the estimation of molecular diversity, they allow one to measure continuous distances and to employ similarity measures (coefficients) which are metrics, contrary to, e.g., fingerprints for which the Jaccard (Tanimoto) coefficient of similarity should be used which is not a metric. Along with other descriptors, such as fingerprints and pharmacophores, TIs can be used for a rapid and efficient virtual screening of databases of hundreds of thousands of compounds. Nevertheless, TIs, being global descriptors of the molecular graph,[17−21] they do not contain explicit information regarding the number of functional groups, pharmacophores, volume, surface area, interatomic distances, charge distribution, orbital energy, or electrostatic potential. But when the size of a compound collection is reduced to a more manageable number, more complex geometric and quantum descriptors may be applied for refining lead selection and optimization. From the large number of TIs used in quantitative structure−activity/structure−property relationships (QSAR/QSPR) models only a few have been tested in molecular diversity studies. The main reasons for this are that they are not available in commercial software packages and that many TIs were initially defined only for hydrocarbons, resulting in a lack of the proper mathematical equations and heteroatom parameters to compute them for any organic compound.

Selection of Descriptors. It is technically impossible to evaluate computationally all of the possible multiparameter correlations of the wealth of descriptors proposed and employed thus far. Furthermore, different descriptors frequently represent, more or less, similar aspects of the molecular structure, and, in this case, they are not able to separate and identify subtle differences in structural features that contribute to the molecular property or activity of interest.[22] To establish the relevance and utility of the various descriptors for data partition in diversity and QSAR/QSPR studies, the reduction of the number of descriptors is often necessary, i.e., the collapse of the dimensions of the chemical space into a space with fewer dimensions, in which case the dimensions refer to the attribute(s) or descriptor(s) that describe a molecular property. This helps to both avoid redundancy from highly correlated descriptors and manage more efficiently the data in terms of computation resources and intuitive perception of the chemical space. Besides, the mere uncritical use of correlated variables to compute a measure of similarity is essentially an implicit weighting of these variables. If, e.g., three variables are used which are highly correlated, the effect is the same as using only one of these variables that has a relative weight three times greater than any other variable. However, rigorous selection of the uniquely best parameters from a large descriptor pool remains an, as yet, unsolved issue, and none of the existing procedures can guarantee the optimal solution.[23] For continuous numerical descriptors the process of the reduction of the number of descriptors usually involves either, or both, of the following approaches:

(i) Construction of the matrix which contains the correlation coefficients that describe the intercorrelations of the continuous values of the descriptors.[24−26] Hence, on the basis of the correlation coefficients, descriptors which have as low as possible correlation coefficients, i.e., that have no (or very little) linear intercorrelations between them, can be selected from the original pool of descriptors.

(ii) Principal component analysis (PCA), or factor analysis (FA) techniques.[27] PCA and FA are statistical tools which build a model by linearly projecting the original variables into much less numerous, more meaningful, and easily manipulated variables. This is achieved by detecting coherent subsets of variables which are independent from each other. Then, the variables within each of these subsets are combined into principal components or factors (latent variables). The mathematical difference is that PCA assumes all variance is common, with all unique factors set to zero, whereas FA considers that there is some unique variance, which is determined by the FA model employed.[27]

Various approaches for the investigation of similarity and pairwise relationships among molecular graph descriptors have been proposed in the literature.[28−34] In one of these studies, a group of 200 TIs, including the Wiener index, connectivity indices, frequency of path lengths of varying size, Balaban index *J*, triplet indices, information theoretic indices defined by Bonchev and Trinajstic, as well as those defined by Basak on the neighborhood complexity of hydrogen-containing molecular graphs were computed for two datasets consisting of 139 hydrocarbons and 1037 diverse chemicals.[34] Then, clustering of these descriptors revealed 14 clusters for the hydrocarbon set and 18 clusters for the diverse chemicals set.

In a previous work we addressed the issue of noncorrelation (quasi-orthogonality) and descriptor selection of novel for molecular diversity studies, graph-theoretical descriptors by means of a heuristic approach.[35] In the present work we further examine whether the same descriptors form "natural groupings" or classes of descriptors. In the formation of such classes it is logical to assume that it is an outcome of the degree of intercorrelation of the descriptor values and their degeneracy with respect to the molecular information they explain. We classified the descriptors into groups by means of cluster analysis (CA) methods, so that descriptors in one group are "similar" to each other, whereas descriptors in different groups are "different" or "diverse" from each other. Since we deal with structural descriptors, similarity and diversity refer to the discriminating power of the different descriptors to account for different structural features/properties in the molecule. Adapting the central dogma of molecular diversity, the similar property principle, descriptors which belong to the same cluster should be expected to account for similar areas of the "chemical space", and descriptors in different clusters should explain different aspects of this space. An additional objective was to evaluate the relative advantages and shortcomings of several different CA methods, to identify the methods which perform well for this specific goal, that is, that do not impose an inappropriate structure on the data and can serve as a good

indication of the diversity and/or the similarity of a dataset. Essentially the idea was to compare observed and predicted closeness, where the observed corresponded to the correlations among the descriptors depicted by their correlation coefficients and the predicted to the cluster appurtenance of the variables.

Strategy. The strategy followed for the discovery of groupings of descriptors was an adaptation of the straightforward technique of cluster validation referred to as *replication*.[36] Generally, replication involves the estimation of the degree that a cluster solution is repeated across a series of data sets. If a cluster solution is repeatedly discovered across different samples from the same general population or different methods using the same sample, it is plausible to conclude that this solution has some generality. A cluster solution that is not stable is unlikely to have general utility. Replication is essentially a test of the internal consistency (i.e., the replicability) of a clustering solution. More specifically, to discover the real structure in the data (groupings of descriptors), we subjected our dataset of molecular descriptors to seven different CA methods and searched to what extent the different descriptors invariably appear together in the same clusters. In other words, we sought the degree of replicability of the clustering solutions across different methods applied on the same sample. Thus, we were interested in exploiting to the greatest extent the subtleties, merits, and inherent biases and deficiencies of the different algorithms in identifying certain types of clusters, instead of fully relying on the validity of a given method. Following this approach, the resulting overall solution across all different methods becomes more reliable and robust. The uniqueness of the clustering approach described here, at least for the field of molecular diversity, stems from the following:

(1) The dataset consisted of molecules described in terms of the molecular descriptors chosen. However, the clustering treated exactly these variables as objects.

(2) The metric employed was the determination coefficient (DEC), i.e., the squared correlation coefficient calculated from the correlation matrix, meaning that the N objects (variables) were each defined on the $N - 1$-dimensional space of the other $N - 1$ variables, according to their relative "closeness", or similarity (or equivalently "distance" or diversity), as this was expressed by means of their correlation coefficients, i.e., the $(N^2 - N)/2$ unique nondiagonal values of the correlation matrix. It must be noted here that, the correlation coefficients are not additive, because the value of the correlation coefficient is not a linear function of the magnitude of the relation between the variables. To average correlations, one has to first convert them into additive measures, such as the DEC.

(3) The resulting structure of the dataset (groups of descriptors) emerged by considering the clustering solutions across different CA methods and on the same level of clustering.

Cluster Analysis. A major class of the pattern recognition techniques which aim at identifying regularities and similarities (i.e., patterns) that exist in the data is CA. CA generically refers to different multivariate methods designed to create homogeneous groups of objects (or cases, or entities) called clusters.[37] Hence, CA classification places objects into more or less homogeneous groups, so that the relationships both within and between groups are revealed. Therefore, CA may also be viewed as a data reduction technique, which groups together either variables or objects based on similar data characteristics. CA is generally considered heuristic in nature, since it requires the user to make decisions related to the calculation or interpretation of clusters, decisions which may have a strong influence on the results of the obtained classification. The most important difference between performing CA and FA for classifying variables is the linear operation of the latter, which for the most part, is not the case in CA. The central question, in any CA method, is whether the cluster structure is meaningful in some way and not arbitrary, or artificial. To do so, the clustering techniques attempt to have more in common within groups than across groups, through either minimization of an objective function all along the clustering process or optimization of a property of the clusters being formed. CA has been a useful tool in chemical information for the prediction of molecular properties and for similarity studies in large databases.[22,38−41]

METHODS

Derivation of the Correlations of Descriptors. The correlation coefficients were calculated starting with a training sample of molecules from the AIDS Database of the Developmental Therapeutics Program of the National Cancer Institute.[42] The entire database contains 32 110 molecules with molecular weight ranging from 26 to 2839. The training sample was composed of 2000 molecules randomly chosen from the pool of the molecules in the entire database which contained only H, C, N, O, P, S, and halogen atoms and had molecular weight less than 660, a range of molecular weight in which fell 95% of the molecules.

Graph-Theoretical Descriptors Used. A set of 240 unique graph-theoretical descriptors was calculated, using various operators and molecular matrices, as well as six different weighting schemes.

Molecular Matrices. The large majority of the topological indices proposed in the literature were derived from the adjacency matrix **A** and the distance matrix **D**. However, the recently introduced reciprocal distance matrix[43−46] **RD** was found to be the source of valuable structural descriptors for structure−property models.[33,47] In this paper the graph descriptors are computed from these three molecular matrices: adjacency **A**, distance **D**, and reciprocal distance **RD** matrices.

Weighting Schemes. In the chemical graph theory, an organic compound containing heteroatoms and/or multiple bonds can be represented as a vertex- and edge-weighted molecular graph.[20] To compute graph structural descriptors, several weighting schemes were proposed by using the atomic number Z,[48] the atomic radius, and electronegativity.[47,49] The molecular descriptors used in this paper were computed with six weighting schemes w, namely, the atomic number Z, the atomic mass A, the hydrogen-augmented atomic mass AH, the atomic polarizability P, the atomic electronegativity E, and the atomic radius R.[50]

The theory and calculation of these descriptors has been described in detail elsewhere.[35] Briefly, they included the following: (1) the molecular weight, **MW**; (2) the Kier and Hall connectivity indices ${}^0\chi^v$, ${}^1\chi^v$, ${}^2\chi^v$, ${}^3\chi_p{}^v$, ${}^3\chi_c{}^v$;[17,18] (3) the **Chi** indices[51] ^{0}**Chi**(**VSD**,w), ^{1}**Chi**(**VSD**,w), ^{2}**Chi**(**VSD**,w), ^{3}**Chi**(**VSD**,w)$_p$, ^{3}**Chi**(**VSD**,w)$_c$, where the vertex invariant **VSD** is the valency **Val**,[52,53] the distance sum **DS**,[19,20] and

the reciprocal distance sum **RDS**;[54] (4) the Wiener indices computed with the Wiener operator **Wi**(**M**^{p},*w*) from the *p*th power of the molecular matrix **M**, with *p* between 1 and 5;[55] (5) the hyper-Wiener indices computed with the hyper-Wiener operator **HyWi**(**M**,*w*);[51,56] (6) the spectral operators **MinSp**(**M**,*w*) and **MaxSp**(**M**,*w*);[51,56] and (7) the Hosoya indices computed with the Hosoya operator **Ho**(**A**,*w*).[52−57]

CA Methods. The CA methods are broadly divided into hierarchical and nonhierarchical:

(A) Hierarchical Clustering (HC) Methods, depending on the criteria of each algorithm, optimize the pathway along which the structure in data is sought. HC involves building a hierarchical classification of the objects in a dataset by a series of either binary fusions (agglomerations) or divisions. A single partition, i.e., a cut across the hierarchy, can then be taken at any level to give the desired number of clusters. In our case we only dealt with HC agglomerative procedures. Given that these methods generate a complete hierarchy of groupings, they are much more demanding in computational resources, with a time complexity from O(N^2) to O(N^3), where *N* is the number of objects to be clustered. The differences of the various algorithms lie in how they group together (link) objects, or progressively built clusters, based on how the closeness (similarity) between them is measured. The methods applied here were:

(I) Single Linkage (or Nearest Neighbor Linkage).[58] The consecutive agglomerations are based on measuring the distance to the nearest object in a group.

(II) Complete Linkage (or Furthest Neighbor Linkage).[59] The linkage rule is still based on smallest distance, but the distances between clusters are calculated on the basis of the furthest objects.

(III) Average Linkage.[37,59] Contrary to the previous two methods, in which the linkages are both based upon single data values within groups, the average linkage is based upon the average distance from all objects in a group. Thus, it averages all distances between objects in different clusters to decide how far apart they are.

(IV) Ward's Method.[60,61] It is a method in which the variance of groups is assessed during the clustering process, and the group which experiences the smallest increase in variance with the iterative inclusion of an object will receive that object.

(B) Nonhierarchical (NHC) Methods attempt to optimize a property of similarity in the groups being formed. NHC methods generally perform a single partition of the dataset, thus producing a classification of (usually) nonoverlapping groups without any hierarchical relationships among them. NHC methods have a time complexity from O(*N*) to O(N^2). The NHC methods employed in this study were as follows:

(I) Jarvis−Patrick (or Nearest Neighbor) Method.[62] It places the objects in the same cluster as a (user-defined) number of their nearest neighbors (the most similar to them other objects), which is a list of a fixed cardinality.

(II) Variable Jarvis−Patrick Method.[63] It is similar to the previous method with the difference that the nearest neighbor lists for each object are of variable length, depending on the value (percentage) of a user-defined threshold similarity.

(III) Relocation, or Iterative Partitioning, or *k*-Means Method.[64] It uses arbitrary objects as cluster "seeds" to begin clustering and then iteratively reassigns the objects into groups, in the search for the "best" division of the objects into different groups so that the within-cluster variance is minimized.

Choice of Clustering Level. Given that CA aims at creating homogeneous groups, it is logical to expect that one of the main issues involved is the determination of the actual number of groups present in the structure that has been recovered. So, e.g., the nested treelike structure of the HC methods (depicted visually by a dendrogram) suggests that many different levels of groupings may be perceived in the data, and the most staggering question is at what level to "cut" (prune) the resulted hierarchy in order to derive the optimal number of groups. Furthermore, the iterative relocation NHC methods require the user to specify the number of clusters in the data prior to the analysis. The same applies to the nearest neighbor NHC methods, as the choice of the parameters by the user influences the number of clusters to be obtained. An additional complication is the fact that CA is computationally demanding. There are $k^n/k!$ ways to assign *N* objects in *k* clusters, and choosing the optimal solution among these is an NP-complete problem. The principle of optimality applied to dynamic programming approaches does not hold for clustering, because the optimal clustering for a subregion of space is likely not to be part of the optimal global clustering. One reason for the complexity of the problem is the interaction between each decision; e.g., changing the assignment of one object to a cluster will change the centers of both clusters, thus affecting the merits of other decisions. Unfortunately, the determination of the "true number" of clusters is a fundamental and, as yet, unresolved issue in CA. The two most important reasons for the little progress is as follows:

(i) The lack of a suitable null hypothesis, largely due to the lack of a consistent and comprehensive definition of what a cluster is, both in terms of structure and content. The negation, i.e., "absence of structure" in a data set as an alternative null hypothesis, is far from clear, and it is not obvious what types of tests could be devised to determine if structure is, or not, present. It should be also added here that, by definition, many CA methods will inevitably create a structure, even artificially. This phenomenon is especially true for the agglomerative hierarchical methods which, with no exception, for *N* initial objects will start with *N* clusters and end up with one, going through a procedure which essentially consists of discovering the best path through the data which will satisfy this. Most of the null hypotheses formulated thus far are limited in scope and not readily applicable to real-world data.[65]

(ii) The complex nature of multivariate sampling distributions. Equally intractable is the problem of the mixture of potentially complex multivariate sampling distributions in the analysis of real-world data. Due to that CA is, by definition, multivariate, in contrast to other classification techniques such as decision trees, which are univariate at each split; clusters are formed in terms of similarities (or distances) considering all of the variables. Although many of the aspects of multivariate normal distributions are well-understood, it is reasonable to expect that many real-world data behave differently. In addition, many datasets may be composed of complex mixtures of different multivariate sampling distributions of unknown structures. Since there is no established statistical theory to unravel those mixtures, it

is probably unreasonable to assume that formal tests of clustering tendency are likely to be easily developed.

To test the structure that really exists in the data, as this emerges across different CA methods, we chose to utilize the most plausible and straightforward approach, that is, to examine on the same level of clustering all seven methods employed. By the very nature of the different methods, only the outcome of an HC method can be examined for optimal clustering level as it is unique for a given dataset (a hierarchy), irrespective of external parameters. Only the intrinsic criteria of the algorithm are important as to how it goes through the data in order to generate the final hierarchy. For the NHC methods the user has to predefine either the number of clusters to be obtained (relocation method) or a set of parameters which have to do with the property to be searched for by the method, that is the stringency of how similarity, or distance, is perceived (nearest neighbor methods). The approach followed here was to examine the clustering levels and prune at the one locally optimal level which has the largest distance to its neighboring levels, that is the largest discontinuity in mean cluster separation as the number of clusters changes. The inflection point at which the slope of the curve of this plot greatly changes corresponds to that level which has the largest distance to its neighboring levels. This change in slope signifies that clustering beyond this level does not produce further large benefits in terms of the clustering of the objects. However, this inflection point does not always correspond to a single point on the curve which describes the clustering procedure. Instead, it may extend over a relatively wide area of clustering levels. As this was the case in the present work, the choice of a common level of reference was necessarily done on the basis of a careful analysis of the obtained hierarchies, as it is described later in Results.

Comparison of the Results of CA Methods. A clear numerical description of the behavior of different CA results may be done by means of calculating the clustering entropy (CE) and the effective number of clusters (ENC) represented.[66,67] In information theoretical terms,[68] variety means multiplicity of distinctions. Increased variety corresponds to increased uncertainty about the outcome of a process. When the uncertainty is relieved by the occurrence of one of the possibilities, then this results in gain of information. Reduction in the quantity of variety is exactly the process of selection. In this case, selection is the reduction of the number of descriptors by clustering them and being able to extract only one member of each cluster which represents the chemical information content conveyed by all the descriptors in its cluster. Entropy refers to the number of distinctions, and CE shows how well-balanced is the data partition among the q clusters. First, for each cluster i we calculate the fraction f_i of the total objects N found in cluster i ($f_i = n_i/N$, where n_i refers to the objects included in cluster i). Then, CE was computed using the following equation:

$$\mathrm{CE} = -\sum_{i=1}^{q} f_i \ln_2 f_i \qquad (1)$$

This expression is based on the information theory[68] according to which if a set of objects (here descriptors) can be classified into q types (here clusters), and if the type (cluster) i has f_i a priori probability of appearance (here substituted by the fraction of objects included in this cluster), then CE is the expected number of bits of information conveyed when we know the type a previously unidentified object belongs to. The greater the number of types (in this case, clusters), and the more uniform the composition, the greater the variability. The ENC is computed by means of the equation

$$\mathrm{ENC} = 2^{\mathrm{CE}} \qquad (2)$$

where ENC gives the number of clusters that would give the observed value of CE if all clusters occurred with equal frequency or were equally populated. If q' of the clusters are large and of approximately equal size and the rest are much smaller, then ENC tends to approach the value of q'.

RESULTS

Common Level of Reference. As discussed above, the inflection point, which signifies that clustering beyond this level does not produce further large benefits in terms of the clustering of the objects, does not always correspond to a single point on the curve which describes the progress of a HC method. Our results showed that for the different HC methods the "best clustering" region, including singleton clusters, fell around 25 and 44 clusters for complete linkage, around 24 and more than 50 clusters for average linkage, around 28 and 56 for single linkage, and around 28 and more than 40 clusters for Ward's method. Given this relative discrepancy, the approach followed was the following: To assess the most significant number of general groups in the data, the idea was to find the minimum number of clusters which well-separates the most orthogonal descriptors.[35] Due to the very nature of the different algorithms, for the 3 out of these 4 methods, with the exception of Ward's method, the 5 more strongly uncorrelated objects (descriptors) were found to belong to different clusters very early in the clustering process. With only 10 clusters they all belonged to different clusters, thus indicating that these methods generally tend to produce clustering results which balance both good separation and representativity in that different clusters do represent different areas of the multivariate space. Ward's method, on the contrary, following the minimum variance criterion, tends to separate strong outliers, namely, the objects which are quite different compared to the majority of the others, and keep them together and apart from the already formed clusters until later in the clustering process. For example, 5 of the most orthogonal descriptors in the entire list (considering the mean DEC of all descriptors), i.e., **Ho**(**A**,*P*), **Ho**(**A**,*R*), ^{3}**Chi**(**RDS**,*P*)$_c$, **MinSp**(**RD**,*E*), and **MinSp**(**RD**,*P*), are kept together, although they are rather different from each other, because they share as commonality among them the fact that they are, collectively, too different from the rest of the descriptors. This of course does separate well outlier and nonoutlier observations. However, the problem arises from the following issue: One of the major uses of CA is to be able to extract one (or just a very few, depending on the sample and study in question) object from each cluster, which would hopefully represent well the properties of the entirety of its cluster, that is, its fellow members. If one were to choose a clustering level which, on one hand, does separate well the outliers but, on the other hand, keeps these outliers together, the choice of a single

object from this cluster of outliers obviously does not guarantee that all of the extremities of the multidimensional space are covered. Certainly, this applies to the case in which the purpose of CA is to sample well the entirety of the initial space, rather than the, say, most populated areas. If one is satisfied with the balance that offers the selection of only one (or a few) outlier observations, in addition to those from the most densely populated areas, then this does not pose any serious problems. In our case, the task was not to sample the entire space but to discover the "real structure" in data, that is, homogeneities among the members of the emerging groups, with the highest degree of confidence. Therefore a good separation among the groups was desirable, and the outlier descriptors should exactly reflect this structure of data. In other words, the strong outliers should belong to distinct groups, so that they could be able to cluster with some of the observations with which they might share commonalties (proximity within the framework of the data structure), or remain as singleton clusters. Given the above-mentioned characteristic of the Ward's method, we found the clustering level at which this method separated these outlier descriptors in different clusters. This occurred at the level of 28 clusters. Therefore, 28 clusters were chosen as the "good" clustering level for uncovering the data structure, and the same number of clusters was obtained with the rest of the methods in order to reference the resulting structure on a common clustering basis. The only difference was the Jarvis−Patrick method with which 29 clusters were obtained.

Discussion on the CA Methods. Most comparisons of clustering methods, thus far, have been based upon the evaluation of how well different methods recover the structure of a dataset with known structure. Thus, standard datasets have been developed for the comparison of different clustering methods in terms of their ability for predicting properties. The most well-known such example is the Iris dataset.[69] In our case, the "known structure" was the result of the intercorrelations of the different descriptors as they are revealed by their correlation coefficients. The sizes of the clusters obtained by applying the seven different CA methods are shown in Table 1, whereas the values of CE and ENC calculated for each method, and upon which the comparison of these methods was based, are shown in Table 2.

HC Methods. (A) Single Linkage. It tends to find long, stringy clusters, a phenomenon known as chaining. This can be verified by considering the results included in Table 1, in which we see that the most populated cluster contains 165 out of totally 240 descriptors, namely, 68.75% of the entire pool. The second most-populated cluster includes 24 members, and the third one has 11 members. Hence, these three clusters contain in all 200, or 83.33% of the entire collection of descriptors. In contrast, of the rest of the 25 clusters, 5 contain three-members, 5 clusters contain only 2 members, and 15 clusters contain only 1 member. By applying the clustering results obtained with single linkage, we get CE = 2.085 and ENC = 4.243. We see that although the clusters are 28, the effective number of clusters is almost 4. This is due to that the new clustered space is composed of three large clusters and a further (fractional) dimension that collectively encompasses the remaining much smaller ones. For this, single linkage is considered to be a space-contracting method, meaning that as the clustering procedure advances one, or a few large clusters have been formed, and either the remaining objects are added, one by one, to these large cluster(s) or they are included in very small clusters. Thus, single linkage gave the single cluster with the largest size (165 members), as well as the highest number of singleton and doubleton clusters (15 singletons and 5 doubletons). This behavior prohibits, in many cases, determination how many clusters really exist in the data. Or, to put it otherwise, it is not always guaranteed that the generated solution accurately recovers the overall structure of the data. This general behavior of the algorithm results, in that when new points are encountered in the space, they tend to join already existing groups rather than be used to start new clusters, and as a group gradually grows it becomes more similar to other groups. Due to the specific clustering criterion, the "properties" of the growing clusters (such as size or diameter) are ignored; the algorithm, in principle, cannot distinguish between two well-populated and discrete clusters connected by a chain of outlying objects (outliers). The single linkage algorithm may combine such clusters

Table 1. Sizes of the Obtained Clusters at the 28-Cluster Level for the Different CA Methods Employed (29-Cluster Level for the Jarvis−Patrick Method)[a]

cluster no.	av linkage	complete linkage	single linkage	Ward's	Jarvis−Patrick	variable Jarvis−Patrick	*k*-means (relocation)
1	37	24	165	24	52	150	21
2	30	23	24	18	27	24	19
3	22	16	11	16	24	15	18
4	22	16	3	14	22	13	17
5	21	15	3	13	18	5	14
6	16	15	3	13	12	3	13
7	16	14	3	12	11	3	13
8	15	13	3	12	7	3	11
9	13	12	2	11	6	3	11
10	11	11	2	11	6	3	10
11	6	11	2	11	6	1	9
12	4	11	2	10	6	1	9
13	3	10	2	10	5	1	8
14	3	8	1	9	5	1	7
15	3	8	1	7	5	1	7
16	3	6	1	7	4	1	6
17	2	4	1	7	4	1	6
18	2	3	1	6	3	1	6
19	2	3	1	5	3	1	5
20	1	3	1	5	2	1	5
21	1	3	1	5	2	1	4
22	1	2	1	3	2	1	4
23	1	2	1	3	2	1	3
24	1	2	1	2	1	1	3
25	1	2	1	2	1	1	3
26	1	1	1	2	1	1	3
27	1	1	1	1	1	1	3
28	1	1	1	1	1	1	2
29	0	0	0	0	1	0	0

[a] The ranking of clusters depends only on their cardinality.

Table 2. Values of the Clustering Entropy and of the Effective Number of Clusters on a Common Level of Clustering: 28 Clusters for the Six Methods and 29-Cluster Level for the Jarvis−Patrick

CA method	clustering entropy	effective no. of clusters
average linkage	3.950	15.455
complete linkage	4.367	20.635
single linkage	2.085	4.243
Ward's	4.496	22.565
Jarvis−Patrick	3.994	15.934
variable Jarvis−Patrick	2.338	5.056
k-means (relocation)	4.539	23.247

along with their outliers into a single large cluster. This can happen even quite early in the process, thereby distorting the overall results.[70,71] For example, we observe in Table 3 that descriptors **HyWi**(**A**,*A*), **HyWi**(**A**,*E*), **HyWi**(**A**,*AH*), **HyWi**(**A**,*R*) and **HyWi**(**A**,*Z*) form a well-defined cluster in 5 out of the 7 clustering methods, in which **HyWi**(**A**,*P*) does not belong. Moreover, descriptors **HyWi**(**D**,*A*), **HyWi**(**D**,*E*), **HyWi**(**D**,*AH*), **HyWi**(**D**,*P*), **HyWi**(**D**,*R*), and **HyWi**(**D**,*Z*) all belong to another well-defined cluster according to all of the methods employed. However, single linkage, due to its rather "loose" clustering criterion, cannot distinguish these clusters and includes all 12 of these descriptors to the very same cluster. Another characteristic example is provided by inspection of all 84 descriptors derived with the **Wi** operator. With the exception of **Wi**(**A**,^{4}*P*), they all belong to the same cluster according to single linkage, which is not the case in any of the other methods.

Concerning the descriptors with the six lowest overall mean DEC values [**Ho**(**A**,*P*), ^{3}**Chi**(**RDS**,*P*)$_c$, **MinSp**(**RD**,*E*), **MinSp**(**RD**,*R*), **MinSp**(**RD**,*P*), **Ho**(**A**,*R*)], as well as the one with the ninth lowest value [^{2}**Chi**(**RDS**,*P*)], they are all included in singleton clusters, according to the single linkage clustering results. The seventh lowest such value, which has descriptor **MinSp**(**A**,*P*), is included in a doubleton cluster with **MaxSp**(**A**,*P*), the descriptor with the seventeenth lowest such value, since they share a pairwise DEC value of 0.830 023. Furthermore, ^{3}**Chi**(**RDS**,*Z*)$_c$, with the eighth lowest mean DEC value, shares the same cluster with ^{3}**Chi**(**RDS**,*E*)$_c$ and ^{3}**Chi**(**RDS**,*R*)$_c$, which have the fifteenth and fourteenth lowest such values, with pairwise DEC 0.769 910 and 0. 769 101, respectively. Interestingly, the descriptors with the tenth [**MinSp**(**RD**,*A*)], eleventh [**MinSp**(**RD**,*AH*)], and twelfth [**MinSp**(**RD**,*Z*)] lowest values of mean DEC belong to the same cluster. The above results show that the extreme "outlier" observations are generally treated as such by the algorithm.

(B) Complete Linkage. As is the case with single linkage clustering, the cluster structure (size and diameter) is again ignored, since the comparisons are based on individual objects and not on entire clusters. However, this algorithm attempts to overcome the problems of chaining by following a more rigorous rule (the most distant neighbor criterion) than the one embodied in single linkage, and, therefore, it tends to find many clusters that are generally very compact, roughly of equivalent size, and hyperspherical (globular) composed of highly similar objects. It is characteristic of that from the three-linkage HC methods complete linkage gives the lowest number of very small clusters. In particular, it produced only three singletons and four doubletons. It follows that the main advantage of complete linkage is that it guarantees a common, maximal internal dissimilarity (diameter) for all clusters, thus producing compact clusters.[72] It is, therefore, a space-dilating method; i.e., as a cluster grows, it tends to become more dissimilar to others. This minimizes intracluster dissimilarities and may be perceived as similar to a minimum-variance clustering. So, we see from the results shown in Table 1 that there are 12 clusters that contain up to 4 descriptors and which include the 27 most outlier descriptors. There are only 3 clusters with 6 to 8 descriptors, 11 clusters with 10 to 16 descriptors, and the remaining 2 clusters with more than 20 descriptors. The result of the above is that by selecting objects from different clusters, we get a more representative sampling of the original space than with single linkage. Again, the most characteristic example comes from the descriptors derived with the **Wi** operator. Thus, complete linkage includes these 84 descriptors in 15 different clusters instead of, essentially, one for single linkage. Further examples can be easily retrieved by inspection of Table 3, especially concerning the descriptors derived with the **Chi** operator.

Of the descriptors with the lowest mean DEC values only the ones with the first [**Ho**(**A**,*P*)] as well as the third such value [**MinSp**(**RD**,*E*)] form singleton clusters. However, the descriptors with the second [^{3}**Chi**(**RDS**,*P*)$_c$,] and the ninth lowest values [^{2}**Chi**(**RDS**,*P*)] form together a doubleton cluster, as is the case for the descriptors ranked fourth [**MinSp**(**RD**,*R*)] and fifth [**MinSp**(**RD**,*P*)] in the lowest mean DEC list. The sixth such descriptor [**Ho**(**A**,*R*)] forms a doubleton with **Ho**(**A**,*AH*), whereas the seventh [**MinSp**(**A**,*P*)] belongs to a cluster with three other descriptors, whereas the eighth [^{3}**Chi**(**RDS**,*Z*)$_c$] and tenth [**MinSp**(**RD**,*A*)] form exactly the same clusters as previously in the case of single linkage. These findings show that the extreme "outlier" observations are less frequently treated as singleton as compared to single linkage. However, a selection from the various small clusters gives a representative selection of the most distant objects.

As a consequence of the above observations, the value of CE was found equal to 4.367 and that of the ENC equal to 20.635, the latter being much closer to the actual number of clusters obtained (28) as compared to single linkage.

(C) Average Linkage. It generates results which lie between those obtained with single linkage and complete linkage. So, the results of Table 1 show that there are two large clusters with 30 and 37 descriptors, one cluster with 21 descriptors, two with 22 members and 5 with 11 to 16 members, two with 6 and 4-members, and, finally, 16 clusters including only 1 to 3-members (the most outlier descriptors). Thus, the method can be considered more balanced, as it is efficient when the objects form natural distinct clumps, as is the case, e.g., for the medium-sized clusters obtained, but it also performs quite well with elongated, large clusters. As a consequence, it gave 12 very small clusters (9 singletons and 3 doubletons) in total, a number which is much closer to that for complete linkage (10), rather than for single linkage (20).

It is evident that since the distances are calculated on the basis of "average distances" between clusters (that is, by considering all the objects in the clusters), the structure and properties of the clusters are indeed considered in the clustering procedure. Thus, average linkage is supposed to be much less affected by the shape of the clusters and is thought to be more capable of detecting "natural" clusters or groupings.[71] In concert with the above, the numerical treatment of the results of average linkage gave CE = 3.950 and ENC = 15.455, values which are between those obtained with single and complete linkage.

Considering the descriptors in the list of the lowest mean DEC values, the ones with the first [**Ho**(**A**,*P*)], the third [**MinSp**(**RD**,*E*)], the fourth [**MinSp**(**RD**,*R*)], the fifth [**MinSp**(**RD**,*P*)], and the sixth such values [**Ho**(**A**,*R*)] form singleton clusters. Similarly to complete linkage, the descriptors with the second [^{3}**Chi**(**RDS**,*P*)$_c$] and the ninth lowest values [^{2}**Chi**(**RDS**,*P*)] form together a doubleton cluster, and

the one with the seventh such value [**MinSp**(**A**,*P*)] belongs to a cluster with two other descriptors. The eighth such descriptor [^{3}**Chi**(**RDS**,*Z*)$_c$] forms nearly the same cluster as those previously discussed in the cases of single and complete linkages, the only difference being the inclusion of the descriptor $^3\chi_c$, too. As for the tenth descriptor in the list of the lowest mean DEC values [**MinSp**(**RD**,*A*)] it does form exactly the same cluster as in the cases of both single and complete linkages.

(D) Ward's Method. This method minimizes dispersion within groups and, due to its fusion rule which states that clusters are joined only if the increase in dispersion (variance) is less for that pair of clusters than for any other pair, does, indeed, take into account the properties of the clusters. Also, Ward's method does not necessarily lead to the optimal solution, due to its very nature, which will "force" a clustering, regardless of the inner true structure of the dataset. The method finds well-shaped and -distributed tight hyperspherical (globular) clusters. This can be evidenced by the fact that it generated only five very small clusters, two singletons and three doubletons, and also by viewing the very well distributed branches of the dendrogram which portrays the clustering procedure. This method also has the tendency to find many clusters of nearly equal sizes. In line with the above, by inspection of the sizes of the obtained clusters (Table 1) we can see that there is only one large cluster with more than 20 descriptors (24), 10 clusters with 11−18 members, 10 clusters with 5−10 descriptors, and the rest 7 clusters with no more than 3-members each. These results of Ward's method gave CE = 4.496 and ENC = 22.565, which were the higher values of all HC methods. In particular, the latter was the closest to the number of clusters obtained, namely, 28.

An examination of the descriptors with the lowest mean DEC values shows that only the ones with the first [**Ho**(**A**,*P*)] as well as the third lowest such value [**MinSp**(**RD**,*E*)] form singleton clusters, as is the case for complete linkage. The second such descriptor [^{3}**Chi**(**RDS**,*P*)$_c$] together with the ninth [^{2}**Chi**(**RDS**,*P*)] and the thirteenth [^{3}**Chi**(**RDS**,*P*)$_p$] form together a cluster. Furthermore, the descriptors ranked fourth [**MinSp**(**RD**,*R*)] and fifth [**MinSp**(**RD**,*P*)] in the lowest mean DEC list also form together a doubleton cluster. On the other hand, the sixth descriptor [**Ho**(**A**,*R*)] forms a doubleton with **Ho**(**A**,*AH*). The seventh [**MinSp**(**A**,*P*)] belongs to a cluster with five other descriptors, whereas the eighth [^{3}**Chi**(**RDS**,*Z*)$_c$] belongs to a cluster with four other descriptors. Finally, the descriptor with the tenth lowest value of mean DEC [**MinSp**(**RD**,*A*)] forms a cluster with 4 other descriptors, contrary to what was observed for it in the results of the three previous algorithms.

These findings show that the extreme outlier objects are very rarely treated as singletons in Ward's method, as compared to the three HC linkage algorithms. The method, however, does separate well the outliers from the majority of the rest of the objects. Therefore, one has to sample objects from all clusters in order to guarantee a representative coverage of the entire original multivariate space with respect to both densely and scarcely populated areas.

NHC Methods. (A) Jarvis−Patrick Method. In terms of CE and ENC, the Jarvis−Patrick method gives results which are much similar to the ones obtained with average linkage, i.e., CE = 3.994 and ENC = 15.934. However, the patterns of cluster memberships of the different descriptors are very different, considering the results included in Table 3. For example, the descriptors **Ho**(**A**,*A*), **Ho**(**A**,*E*), and **Ho**(**A**,*Z*) form a single cluster in this case, whereas with average linkage they do belong to the same cluster, which, however, includes 18 more descriptors. The latter form together an 18-member cluster with Jarvis−Patrick (descriptors **Wi**(**D**^{p},*w*) with *p* between 3 and 5 and all 6 weighting schemes *w*). In an additional example, descriptors **Ho**(**A**,*AH*) and **Ho**(**A**,*R*) form a doubleton cluster with Jarvis−Patrick, but two singleton clusters with average linkage.

The Jarvis−Patrick method tends to produce many clusters with only a few members, especially when strict clustering conditions are set or a few very large clusters with very diverse objects when less strict conditions are applied. For example, applying the strict criterion of 5 nearest neighbors and 3 in common, it gives 108 clusters, whereas the more relaxed criterion of 6 nearest neighbor lists and 2 objects in common gives 37 clusters, and the even less strict conditions of 15 nearest neighbors and 6 in common give only 11 clusters (results not shown). The clustering presented here resulted by applying the criterion 7 nearest neighbors and 2 in common. In this case, we can see from Table 1 that there is a very large cluster with 52 descriptors, with three additional clusters still much larger than the majority of the obtained clusters, with 27, 24, and 22 members. Thus, these four out of the totally 29 clusters (13.79%), contain 125 (52.08%) of all the descriptors. However, 22 clusters include between 1 and 7 members, and although they represent more than three-quarters of all clusters (75.86%), they contain less than one-third of all descriptors (30.83%), in total. An additional conclusion which can be derived is that more than one-third of the obtained clusters are either singleton or doubleton clusters (10 in total).

Also, there is no explicit distance consideration between the objects, except in forming the neighbor lists. Hence, relatively dissimilar objects in sparse areas of space can be put together in the same clusters, and large clusters of relatively similar objects can be arbitrarily split in dense areas of space. An example of this is the group of descriptors, **Ho**(**A**,*A*), **Ho**(**A**,*E*), **Ho**(**A**,*Z*), **Wi**(**D**^{4},*A*), **Wi**(**D**,*A*), **Wi**(**D**^{4},*E*), **Wi**(**D**^{5},*E*), **Wi**(**D**^{5},*AH*), **Wi**(**D**^{5},*R*), **Wi**(**D**^{4},*Z*), **Wi**(**D**^{5},*Z*), which belong in all six other methods invariably in the same cluster, but, according to the Jarvis−Patrick method, they are split into two subgroups, one with the first three and another with the remaining eight, in which case this latter cluster also contains 10 more descriptors. Another similar example stems from the examination of the pair of descriptors **HyWi**(**A**,*R*) and **HyWi**(**A**,*Z*) which, being very much correlated with respect to each other with pairwise DEC value 0.946 193, they should normally belong to the same cluster. In fact, this is the case in all CA methods studied with the sole exception of Jarvis−Patrick. On the other hand, considering the descriptors with the lower overall mean DEC values, we can observe in Table 3 the following. The descriptors **MinSp**(**RD**,*E*), **MinSp**(**RD**,*R*), **MinSp**(**RD**,*P*), **MinSp**(**RD**,*A*), **MinSp**(**RD**,*AH*), and **MinSp**(**RD**,*Z*), which have, respectively, the third, fourth, fifth, tenth, eleventh, and twelfth lower mean DEC values of all 240 descriptors, all belong to the same cluster, according to the Jarvis−Patrick method. However, there are only three descriptors among these six which are highly correlated pairwise. Namely, the

Table 3. Cluster Membership of All the Employed Descriptors for Every CA Method at the Common Level of Reference of 28 Clusters (29 for the Jarvis−Patrick Method), As Explained in the Text

descriptor		av	complete	single		Jarvis−Patrick	variable Jarvis−Patrick	
no.	name	linkage	linkage	linkage	Ward's	(7/2)	(25/74)	relocation
1	**Ho**(**A**,*A*)	1	1	1	1	1	1	14
2	**Ho**(**A**,*E*)	1	1	1	1	1	1	14
3	**Ho**(**A**,*AH*)	2	3	25	3	2	2	8
4	**Ho**(**A**,*P*)	4	8	28	4	3	3	8
5	**Ho**(**A**,*R*)	3	3	26	3	2	4	8
6	**Ho**(**A**,*Z*)	1	1	1	1	1	1	14
7	**HyWi**(**A**,*A*)	12	18	1	18	4	1	9
8	**HyWi**(**A**,*E*)	12	18	1	18	4	1	9
9	**HyWi**(**A**,*AH*)	12	18	1	18	5	1	9
10	**HyWi**(**A**,*P*)	23	23	1	23	6	5	24
11	**HyWi**(**A**,*R*)	12	18	1	18	7	1	9
12	**HyWi**(**A**,*Z*)	12	18	1	18	4	1	9
13	**HyWi**(**D**,*A*)	15	22	1	21	8	1	7
14	**HyWi**(**D**,*E*)	15	22	1	21	8	1	7
15	**HyWi**(**D**,*AH*)	15	22	1	21	8	1	7
16	**HyWi**(**D**,*P*)	15	22	1	21	8	1	7
17	**HyWi**(**D**,*R*)	15	22	1	21	8	1	7
18	**HyWi**(**D**,*Z*)	15	22	1	21	8	1	7
19	**HyWi**(**RD**,*A*)	16	25	1	26	9	1	25
20	**HyWi**(**RD**,*E*)	13	20	1	20	10	1	12
21	**HyWi**(**RD**,*AH*)	16	25	1	26	9	1	25
22	**HyWi**(**RD**,*P*)	13	21	1	20	10	1	11
23	**HyWi**(**RD**,*R*)	13	21	1	20	10	1	11
24	**HyWi**(**RD**,*Z*)	16	25	1	26	9	1	25
25	MW	13	20	1	20	10	1	12
26	$^{0}\chi$	13	20	1	19	10	1	12
27	$^{1}\chi$	13	20	1	19	10	1	12
28	$^{2}\chi$	14	28	3	24	11	6	20
29	$^{3}\chi_{p}$	14	28	3	24	11	7	20
30	$^{3}\chi_{c}$	25	13	7	11	12	8	1
31	**MinSp**(**A**,*A*)	26	15	8	9	13	9	23
32	**MaxSp**(**A**,*A*)	26	15	8	9	13	9	23
33	**MinSp**(**A**,*E*)	26	15	8	9	13	9	23
34	**MaxSp**(**A**,*E*)	26	15.	8	9	13	9	23
35	**MinSp**(**A**,*AH*)	26	15	8	9	13	9	23
36	**MaxSp**(**A**,*AH*)	26	15	8	9	13	9	23
37	**MinSp**(**A**,*P*)	5	4	11	7	6	10	8
38	**MaxSp**(**A**,*P*)	5	4	11	7	6	11	8
39	**MinSp**(**A**,*R*)	26	15	12	7	6	11	23
40	**MaxSp**(**A**,*R*)	26	15	12	7	6	9	23
41	**MinSp**(**A**,*Z*)	26	15	8	9	13	9	23
42	**MaxSp**(**A**,*Z*)	26	15	8	9	13	9	23
43	**MinSp**(**D**,*A*)	15	19	1	22	14	1	7
44	**MaxSp**(**D**,*A*)	15	19	1	22	15	1	21
45	**MinSp**(**D**,*E*)	15	19	1	22	14	1	7
46	**MaxSp**(**D**,*E*)	15	19	1	22	15	1	21
47	**MinSp**(**D**,*AH*)	15	19	1	22	14	1	21
48	**MaxSp**(**D**,*AH*)	15	19	1	22	15	1	21
49	**MinSp**(**D**,*P*)	15	19	1	22	14	1	7
50	**MaxSp**(**D**,*P*)	15	19	1	22	15	1	21
51	**MinSp**(**D**,*R*)	15	19	1	22	14	1	7
52	**MaxSp**(**D**,*R*)	15	19	1	22	15	1	21
53	**MinSp**(**D**,*Z*)	15	19	1	22	14	1	7
54	**MaxSp**(**D**,*Z*)	15	19	1	22	15	1	21
55	**MinSp**(**RD**,*A*)	8	9	18	13	16	12	27
56	**MaxSp**(**RD**,*A*)	18	10	1	14	9	1	27
57	**MinSp**(**RD**,*E*)	11	6	27	5	16	13	3
58	**MaxSp**(**RD**,*E*)	17	28	1	27	10	1	6
59	**MinSp**(**RD**,*AH*)	8	9	18	13	16	12	27
60	**MaxSp**(**RD**,*AH*)	18	10	1	14	9	1	27
61	**MinSp**(**RD**,*P*)	9	7	23	6	16	14	3
62	**MaxSp**(**RD**,*P*)	17	26	1	28	10	1	28
63	**MinSp**(**RD**,*R*)	10	7	24	6	16	15	3
64	**MaxSp**(**RD**,*R*)	17	28	1	27	10	1	28
65	**MinSp**(**RD**,*Z*)	8	9	18	13	16	12	27
66	**MaxSp**(**RD**,*Z*)	18	10	1	14	9	1	27
67	**Wi**(**A**^{1},*A*)	12	18	1	18	17	1	9
68	**Wi**(**A**^{2},*A*)	12	18	1	24	18	1	13
69	**Wi**(**A**^{3},*A*)	24	24	1	25	6	16	2
70	**Wi**(**A**^{4},*A*)	24	24	1	25	6	16	2

Table 3. (Continued)

descriptor		av	complete	single		Jarvis−Patrick	variable Jarvis−Patrick	
no.	name	linkage	linkage	linkage	Ward's	(7/2)	(25/74)	relocation
71	**Wi**(**A**^{1},*E*)	12	18	1	18	17	1	9
72	**Wi**(**A**^{2},*E*)	12	18	1	24	18	1	13
73	**Wi**(**A**^{3},*E*)	24	24	1	25	6	16	2
74	**Wi**(**A**^{4},*E*)	24	24	1	25	6	16	2
75	**Wi**(**A**^{1},*AH*)	12	18	1	18	17	1	9
76	**Wi**(**A**^{2},*AH*)	12	18	1	24	18	1	13
77	**Wi**(**A**^{3},*AH*)	24	24	1	25	6	16	2
78	**Wi**(**A**^{4},*AH*)	24	24	1	25	6	16	2
79	**Wi**(**A**^{1},*P*)	12	18	1	18	7	1	9
80	**Wi**(**A**^{2},*P*)	23	23	1	23	6	5	24
81	**Wi**(**A**^{3},*P*)	24	24	1	23	6	17	24
82	**Wi**(**A**^{4},*P*)	5	4	13	7	6	11	24
83	**Wi**(**A**^{1},*R*)	12	18	1	18	17	1	9
84	**Wi**(**A**^{2},*R*)	12	18	1	24	18	1	13
85	**Wi**(**A**^{3},*R*)	24	24	1	23	6	16	2
86	**Wi**(**A**^{4},*R*)	24	24	1	23	6	16	2
87	**Wi**(**A**^{1},*Z*)	12	18	1	18	17	1	9
88	**Wi**(**A**^{2},*Z*)	12	18	1	24	18	1	13
89	**Wi**(**A**^{3},*Z*)	24	24	1	25	6	16	2
90	**Wi**(**A**^{4},*Z*)	24	24	1	25	6	16	2
91	**Wi**(**D**^{1},*A*)	15	19	1	22	19	1	21
92	**Wi**(**D**^{2},*A*)	15	22	1	21	8	1	7
93	**Wi**(**D**^{3},*A*)	1	2	1	2	20	1	14
94	**Wi**(**D**^{4},*A*)	1	1	1	1	20	1	14
95	**Wi**(**D**^{5},*A*)	1	1	1	1	20	1	14
96	**Wi**(**D**^{1},*E*)	15	19	1	22	19	1	21
97	**Wi**(**D**^{2},*E*)	15	22	1	21	8	1	7
98	**Wi**(**D**^{3},*E*)	1	2	1	2	20	1	14
99	**Wi**(**D**^{4},*E*)	1	1	1	1	20	1	14
100	**Wi**(**D**^{5},*E*)	1	1	1	1	20	1	14
101	**Wi**(**D**^{1},AH)	15	19	1	22	21	1	21
102	**Wi**(**D**^{2},AH)	15	22	1	21	8	1	7
103	**Wi**(**D**^{3},AH)	1	2	1	2	20	1	14
104	**Wi**(**D**^{4},AH)	1	2	1	2	20	1	14
105	**Wi**(**D**^{5},*AH*)	1	1	1	1	20	1	14
106	**Wi**(**D**^{1},*P*)	15	19	1	22	22	1	21
107	**Wi**(**D**^{2},*P*)	15	22	1	21	8	1	7
108	**Wi**(**D**^{3},*P*)	1	2	1	2	20	1	14
109	**Wi**(**D**^{4},*P*)	1	2	1	2	20	1	14
110	**Wi**(**D**^{5},*P*)	1	2	1	2	20	1	14
111	**Wi**(**D**^{1},*R*)	15	19	1	22	19	1	21
112	**Wi**(**D**^{2},*R*)	15	22	1	21	8	1	7
113	**Wi**(**D**^{3},*R*)	1	2	1	2	20	1	14
114	**Wi**(**D**^{4},*R*)	1	2	1	2	20	1	14
115	**Wi**(**D**^{5},*R*)	1	1	1	1	20	1	14
116	**Wi**(**D**^{1},*Z*)	15	19	1	22	19	1	21
117	**Wi**(**D**^{2},*Z*)	15	22	1	21	8	1	7
118	**Wi**(**D**^{3},*Z*)	1	2	1	2	20	1	14
119	**Wi**(**D**^{4},*Z*)	1	1	1	1	20	1	14
120	**Wi**(**D**^{5},*Z*)	1	1	1	1	20	1	14
121	**Wi**(**RD**^{1},*A*)	13	21	1	20	10	1	18
122	**Wi**(**RD**^{2},*A*)	16	25	1	26	9	1	25
123	**Wi**(**RD**^{3},*A*)	16	25	1	26	9	1	10
124	**Wi**(**RD**^{4},*A*)	18	27	1	14	9	1	10
125	**Wi**(**RD**^{5},*A*)	18	10	1	14	9	1	27
126	**Wi**(**RD**^{1},*E*)	13	20	1	20	10	1	12
127	**Wi**(**RD**^{2},*E*)	13	21	1	20	10	1	6
128	**Wi**(**RD**^{3},*E*)	13	21	1	27	10	1	6
129	**Wi**(**RD**^{4},*E*)	17	28	1	27	10	1	6
130	**Wi**(**RD**^{5},*E*)	17	28	1	27	10	1	6
131	**Wi**(**RD**^{1},*AH*)	13	21	1	20	10	1	18
132	**Wi**(**RD**^{2},*AH*)	16	25	1	26	9	1	25
133	**Wi**(**RD**^{3},*AH*)	16	25	1	26	9	1	25
134	**Wi**(**RD**^{4},*AH*)	18	27	1	14	9	1	10
135	**Wi**(**RD**^{5},*AH*)	18	10	1	14	9	1	27
136	**Wi**(**RD**^{1},*P*)	13	21	1	20	10	1	11
137	**Wi**(**RD**^{2},*P*)	17	28	1	27	10	1	28
138	**Wi**(**RD**^{3},*P*)	17	28	1	27	10	1	28
139	**Wi**(**RD**^{4},*P*)	17	26	1	28	10	1	28
140	**Wi**(**RD**^{5},*P*)	17	26	1	28	10	1	28
141	**Wi**(**RD**^{1},*R*)	13	20	1	20	10	1	11
142	**Wi**(**RD**^{2},*R*)	13	21	1	20	10	1	11

Table 3. (Continued)

descriptor		av	complete	single		Jarvis−Patrick	variable Jarvis−Patrick	
no.	name	linkage	linkage	linkage	Ward's	(7/2)	(25/74)	relocation
143	**Wi(RD^3,***R***)**	17	28	1	27	0	1	11
144	**Wi(RD^4,***R***)**	17	28	1	27	10	1	28
145	**Wi(RD^5,***R***)**	17	28	1	27	10	1	28
146	**Wi(RD^1,***Z***)**	13	21	1	20	10	1	18
147	**Wi(RD^2,***Z***)**	16	25	1	26	9	1	25
148	**Wi(RD^3,***Z***)**	16	25	1	26	9	1	10
149	**Wi(RD^4,***Z***)**	18	27	1	14	9	1	10
150	**Wi(RD^5,***Z***)**	18	10	1	14	9	1	27
151	**^{0}Chi(DS,***A***)**	21	11	19	12	23	18	22
152	**^{1}Chi(DS,***A***)**	27	16	17	15	24	19	4
153	**^{2}Chi(DS,***A***)**	27	17	17	16	24	19	5
154	**^{3}Chi(DS,***A***)$_p$**	27	17	17	16	24	19	5
155	**^{3}Chi(DS,***A***)$_c$**	28	17	17	17	24	19	15
156	**^{0}Chi(DS,***E***)**	21	11	20	12	23	18	22
157	**^{1}Chi(DS,***E***)**	27	16	17	15	24	19	4
158	**^{2}Chi(DS,***E***)**	27	16	17	15	24	19	26
159	**^{3}Chi(DS,***E***)$_p$**	27	16	17	15	24	19	26
160	**^{3}Chi(DS,***E***)$_c$**	27	17	17	16	24	19	17
161	**^{0}Chi(DS,***AH***)**	21	11	19	12	23	18	22
162	**^{1}Chi(DS,***AH***)**	27	16	17	15	24	19	4
163	**^{2}Chi(DS,***AH***)**	27	17	17	16	24	19	5
164	**^{3}Chi(DS,***AH***)$_p$**	27	17	17	16	24	19	5
165	**^{3}Chi(DS,***AH***)$_c$**	28	17	17	17	24	19	15
166	**^{0}Chi(DS,***P***)**	21	11	20	12	23	18	22
167	**'Chi(DS,***P***)**	27	16	17	15	24	19	4
168	**^{2}Chi(DS,***P***)**	27	16	17	16	24	19	19
169	**^{3}Chi(DS,***P***)**	27	17	17	16	24	19	19
170	**Chi(DS,***P***)$_c$**	27	17	17	16	24	19	15
171	**^{0}Chi(DS,***R***)**	21	11	20	12	23	18	22
172	**^{1}Chi(DS,***R***)**	27	16	17	15	24	19	4
173	**^{2}Chi(DS,***R***)**	27	16	17	15	24	19	26
174	**^{3}Chi(DS,***R***)$_p$**	27	16	17	15	24	19	19
175	**^{3}Chi(DS,***R***)$_c$**	27	17	17	16	24	19	17
176	**^{0}Chi(DS,***Z***)**	21	11	21	12	23	20	22
177	**^{1}Chi(DS,***Z***)**	27	16	17	15	24	19	4
178	**^{2}Chi(DS,***Z***)**	27	16	17	15	24	19	26
179	**^{3}Chi(DS,***Z***)$_p$**	27	16	17	15	24	19	26
180	**^{3}Chi(DS,***Z***)$_c$**	27	17	17	16	24	19	17
181	**^{0}Chi(Val,***A***)**	13	20	1	20	10	1	18
182	**^{1}Chi(Val,***A***)**	13	21	1	20	10	1	18
183	**^{2}Chi(Val,***A***)**	16	27	1	26	9	1	10
184	**^{3}Chi(Val,***A***)$_p$**	16	25	1	26	9	1	25
185	**^{3}Chi(Val,***A***)$_c$**	18	10	4	13	25	1	27
186	**^{0}Chi(Val,***E***)**	13	20	1	20	10	1	12
187	**^{1}Chi(Val,***E***)**	13	20	1	20	10	1	12
188	**^{2}Chi(Val,***E***)**	13	21	1	20	10	1	18
189	**^{3}Chi(Val,***E***)$_p$**	17	28	1	27	10	1	6
190	**^{3}Chi(Val,***E***)$_c$**	19	27	5	11	26	21	10
191	**^{0}Chi(Val,***AH***)**	13	20	1	20	10	1	12
192	**^{1}Chi(Val,***AH***)**	13	21	1	20	10	1	18
193	**^{2}Chi(Val,***AH***)**	16	27	1	26	9	1	10
194	**^{3}Chi(Val,***AH***)$_p$**	16	25	1	26	9	1	25
195	**^{3}Chi(Val,***AH***)$_c$**	18	10	4	13	25	1	27
196	**^{0}Chi(Val,***P***)**	13	20	1	20	10	1	12
197	**^{1}Chi(Val,***P***)**	13	21	1	20	10	1	11
198	**^{2}Chi(Val,***P***)**	17	28	1	27	10	1	28
199	**^{3}Chi(Val,***P***),**	17	28	1	27	10	1	28
200	**^{3}Chi(Val,***P***)$_c$**	19	27	5	11	26	21	10
201	**^{0}Chi(Val,***R***)**	13	20	1	20	10	1	12
202	**^{1}Chi(Val,***R***)**	13	20	1	20	10	1	11
203	**^{2}Chi(Val,***R***)**	13	21	1	20	10	1	11
204	**^{3}Chi(Val,***R***)$_p$**	17	28	1	27	10	1	28
205	**^{3}Chi(Val,***R***)$_c$**	19	27	5	11	26	21	10
206	**^{0}Chi(Val,***Z***)**	13	20	1	19	10	1	12
207	**^{1}Chi(Val,***Z***)**	13	20	1	19	10	1	12
208	**^{2}Chi(Val,***Z***)**	13	20	1	19	10	1	12
209	**^{3}Chi(Val,***Z***)$_p$**	12	20	1	19	27	1	12
210	**^{3}Chi(Val,***Z***)$_c$**	20	13	6	11	26	22	20
211	**^{0}Chi(RDS,***P***)**	12	19	1	18	28	1	9
212	**^{1}Chi(RDS,***P***)**	23	23	2	23	6	5	24
213	**^{2}Chi(RDS,***P***)**	7	5	14	8	6	23	8
214	**^{3}Chi(RDS,***P***)$_p$**	6	4	16	8	6	24	8

Table 3. (Continued)

descriptor		av	complete	single			variable	
no.	name	linkage	linkage	linkage	Ward's	Jarvis−Patrick (7/2)	Jarvis−Patrick (25/74)	relocation
215	^{3}**Chi**(**RDS**,*P*)$_{c}$	7	5	15	8	6	25	8
216	^{0}**Chi**(**RDS**,*A*)	13	20	1	19	10	1	9
217	^{1}**Chi**(**RDS**,*A*)	12	19	1	18	28	1	9
218	^{2}**Chi**(**RDS**,*A*)	24	24	1	25	6	16	2
219	^{3}**Chi**(**RDS**,*A*)$_{p}$	26	15	8	9	13	9	23
220	^{3}**Chi**(**RDS**,*A*)$_{c}$	26	15	9	10	12	9	16
221	^{0}**Chi**(**RDS**,*E*)	13	20	1	19	10	1	12
222	^{1}**Chi**(**RDS**,*E*)	12	19	1	18	28	1	9
223	^{2}**Chi**(**RDS**,*E*)	24	24	1	25	6	16	2
224	^{3}**Chi**(**RDS**,*E*)$_{p}$	26	15	8	9	13	9	23
225	^{3}**Chi**(**RDS**,*E*)$_{c}$	25	14	10	10	12	26	1
226	^{0}**Chi**(**RDS**,*AH*)	13	20	1	19	10	1	9
227	^{1}**Chi**(**RDS**,*AH*)	12	19	1	18	28	1	9
228	^{2}**Chi**(**RDS**,*AH*)	24	24	1	25	6	16	2
229	^{3}**Chi**(**RDS***AH*)$_{p}$	26	15	8	9	13	9	23
230	^{3}**Chi**(**RDS**,*AH*)$_{c}$	26	15	9	10	12	9	16
231	^{0}**Chi**(**RDS**,*R*)	13	20	1	19	10	1	9
232	^{1}**Chi**(**RDS**,*R*)	12	19	1	18	28	1	9
233	^{2}**Chi**(**RDS**,*R*)	24	24	1	23	6	27	24
234	^{3}**Chi**(**RDS**,*R*)$_{p}$	26	15	12	7	6	9	23
235	^{3}**Chi**(**RDS**,*R*)$_{c}$	25	14	10	10	12	26	1
236	^{0}**Chi**(**RDS**,*Z*)	13	20	1	19	10	1	12
237	^{1}**Chi**(**RDS**,*Z*)	13	20	1	19	10	1	12
238	^{2}**Chi**(**RDS**,*Z*)	13	20	1	19	29	1	12
239	^{3}**Chi**(**RDS**,*Z*)$_{p}$	22	12	22	12	23	28	22
240	^{3}**Chi**(**RDS**,*Z*)$_{c}$	25	14	10	10	12	26	1

pair **MinSp**(**RD**,*A*)−**MinSp**(**RD**,*AH*) has DEC 0.997 003, the pair **MinSp**(**RD**,*A*)−**MinSp**(**RD**,*Z*) with DEC value 0.989 414, and the pair **MinSp**(**RD**,*AH*)−**MinSp**(**RD**,*Z*) with DEC value 0.984 772. Most importantly, the first of these six descriptors, **MinSp**(**RD**,*E*), has pairwise DEC values with the other five which are as low as 0.255 250, 0.001 400, 0.050 455, 0.042 649, and 0.058 252, respectively. Nevertheless, it is included in the same cluster with them. Another similar example is the descriptor **MinSp**(**A**,*P*), which is in the same cluster with ^{2}**Chi**(**RDS**,*P*) and ^{3}**Chi**(**RDS**,*P*)$_{c}$, although it has pairwise DEC values with these two descriptors only as low as 0.068 134 and 0.230 065, respectively. However, the latter two descriptors are quite correlated with each other with a pairwise DEC 0.666 145.

Last, this method is very user-dependent, and, therefore, the parameters are generally not transferable from one dataset to another. In addition, several trials and careful analysis of the obtained results are required before the optimal data partition can be obtained.

(B) Variable Jarvis−Patrick Method. The results for this method are CE = 2.338 and ENC = 5.056. Therefore, this method was found to perform very much in the same way as single linkage, producing long stringy clusters (chaining effect) and several clusters with scarce populations. From Table 1 we can see that the most populated cluster contains 150 descriptors and the second most populated one has 24 descriptors, whereas the third and fourth largest clusters have 15 and 13 members, respectively. That is, 84.17% of all the descriptors are included in only 14.29% of the clusters. Of the other clusters, one contains 5 descriptors, 5 have 3 descriptors, and 18 clusters have only one member. Namely, 85.71% of all clusters (24) include only 15.83% of the total number of objects (38 descriptors). It is also interesting to note that the number of singletons produced was the largest of all seven CA methods employed.

Variable Jarvis−Patrick also tends to arbitrarily separate rather similar objects. This is the case, e.g., for the three pairs of descriptors: $^{2}\chi-^{3}\chi_{p}$, **MinSp**(**A**,*P*)−**MaxSp**(**A**,*P*), and **MinSp**(**A**,*R*)−**MaxSp**(**A**,*R*) for which the pairwise DEC values are 0.911 113, 0.830 023, and 0.820 612, respectively. Therefore, it is logical that the descriptors of each pair appear in the same clusters in the CA methods studied. The only exception was the variable Jarvis−Patrick method.

Considering the descriptors with the lowest mean DEC values, the ones with the first [**Ho**(**A**,*P*)], second [^{3}**Chi**(**3 RDS**,*P*)$_{c}$,], third [**MinSp**(**RD**,*E*)], fourth [**MinSp**(**RD**,*R*)], fifth [**MinSp**(**RD**,*P*)], sixth [**Ho**(**A**,*R*)], seventh [**MinSp**(**A**,*P*)], ninth [^{2}**Chi**(**RDS**,*P*)], and thirteenth lowest such values [^{3}**Chi**(**RDS**,*P*)$_{p}$] do form singleton clusters. The descriptor with the eighth lowest value [^{3}**Chi**(**RDS**,*Z*)$_{c}$] forms a cluster with those that have the fourteenth [^{3}**Chi**(**RDS**,*R*)$_{c}$] and fifteenth [^{3}**Chi**(**RDS**,*E*)$_{c}$] such values. Last, the descriptors with the tenth [**MinSp**(**RD**,*A*)], eleventh [**MinSp**(**RD**,*AH*)], and twelfth [**MinSp**(**RD**,*Z*)] lowest mean DEC values cluster together, as they do in all seven CA methods applied.

Furthermore, an obvious disadvantage of this method is that it depends very much on the choice of initial parameters, and in order to achieve the desired clustering level a lot of trials must first be attempted. In this case, the set of the initial parameters applied included a distance threshold (or minimum similarity) equal to 25% of the distance range between the objects, and at least 74% commonality of the shorter nearest−neighbor lists between any two objects, as the minimum requirement in order to include them in the same cluster.

(C) K-Means Relocation Method. The major advantages of this approach are as follows:

(i) Unlike HC methods, which require the calculation and storage of an $N \times N$ similarity matrix for all N objects, and

nearest neighbor NHC which require the calculation of the neighbor lists of every single object in the dataset, this method works directly upon the raw data, making it much easier to handle larger datasets.

(ii) Since the method makes iterative passes through the data, it can compensate for poor initial partitions of the data.

(iii) The method gives a single ranking of clusters that are not nested and, therefore, are not part of a hierarchy, avoiding the problem of HC methods concerning the decision of where to prune the dendrogram.

However, the relocation method suffers from one major limitation. The absolutely correct way to discover the optimal partition of a dataset by means of an iterative method is to form all possible partitions of that data set. Except for the most trivial of problems, this task is computationally intractable. The sampling of only a very small proportion of all possible solutions creates the major drawback of iterative partition clustering methods, that of choosing suboptimal solutions, a problem which is equivalent to locating the local, instead of the global, energy minima in physicochemical or biophysical studies. Unfortunately, there is no objective way to determine if a solution obtained is *globally* or *locally* optimal. This local optimality is pronounced by the following examples. Descriptor **Ho**(**A**,*A*) belongs to the same cluster with 20 other descriptors, with which it has values of DEC between nearly unity [(**Ho**(**A**,*E*) and **Ho**(**A**,*Z*)] and 0.373 37 [**Wi**(**D**^{3},*P*)]. On the other hand, **Ho**(**A**,*P*), which is definitely less correlated with all of the other descriptors, with a mean DEC of 0.007 157 and pairwise DEC against every single descriptor of less than 0.1 is not treated as an outlier object. Instead it belongs to the same cluster with 7 other descriptors. These seven, although generally poorly correlated with the entire pool of descriptors, are somehow correlated with each other. So, **Ho**(**A**,*AH*) is correlated with **Ho**(**A**,*R*) with a DEC = 0.4226. **MinSp**(**A**,*P*) and **MaxSp**(**A**,*P*) are strongly correlated with each other (0.830 02) and relatively (at least 0.23) with ^{2}**Chi**(**RDS**,*P*), ^{3}**Chi**(**RDS**,*P*)$_p$, and ^{3}**Chi**(**RDS**,*P*)$_c$. These effects force them together in the same cluster during the iterative relocation procedure. For **Ho**(**A**,*P*), the best possible solution is to enter a cluster in which belongs a descriptor with which it shares a "minimum" similarity, and this occurs for **MinSp**(**A**,*P*) with which it has a DEC of 0.0220. Therefore, we see that there belong descriptors which altogether may not share strong commonalties, but they form a group by merit of that, first, they are generally quite different compared to the rest so they are forced to cluster somewhere and, second, they have some similarity with at least one member of the cluster (reminiscent of the single linkage rule). The same apply for the cluster which includes descriptors **MinSp**(**RD**,*E*), **MinSp**(**RD**,*P*), and **MinSp**(**RD**,*R*) which are, again, of the most poorly correlated descriptors in the entire set. The first two are almost orthogonal with respect to each other, with a DEC of 0.0014. However, **MinSp**(**RD**,*R*) shares commonalties with both of them, having coefficients 0.2552 and 0.3743 with **MinSp**(**RD**,*E*) and **MinSp**(**RD**,*P*), respectively. This pulls them together to the same (locally optimal) cluster. Another similar example includes the descriptor **MinSp**(**A**,*P*), which is in the same cluster with ^{2}**Chi**(**RDS**,*P*) and ^{3}**Chi**(**RDS**,*P*)$_c$, although it has pairwise DEC values with these two descriptors only as low as 0.068 134 and 0.230 065, respectively. However, the other two descriptors, ^{2}**Chi**(**RDS**,*P*) and ^{3}**Chi**(**RDS**,*P*)$_c$, are quite correlated with each other (pairwise DEC of 0.666 145).

The relocation method performs well for data structured in well-formed hyperspherical clusters.[70,71] This is verified by considering that the calculated values for CE (4.539) and ENC (23.247) were the highest values of all seven methods studied. Furthermore, the results in Table 1 show that the distribution of cluster sizes is much smoother as compared to the other methods. In particular, there are four relatively large clusters containing between 21 and 19 descriptors, whereas three clusters have 13 to 14 descriptors, with another three having 10 to 11 descriptors. Eight include from 6 to 9-members and nine from 3 to 5-members. Finally, there are no singleton clusters, while there is only a single case of a doubleton cluster.

Validation of Clusters of Descriptors. After every CA method was performed, each descriptor was assigned to a certain cluster for every method (results included in Table 3). For each descriptor a seven-digit code was stored, each digit being the cluster it belongs to in each method, the order of which was average linkage, complete linkage, single linkage, Ward's method, Jarvis−Patrick, variable Jarvis−Patrick, and relocation. Then the analysis was performed according to the following steps:

(i) First Step: Construction of Strong Clusters. The clusters which included members that showed the same behavior in terms of what cluster they belonged to, across all 7 different methods, were referred to as *strong* clusters. That is, we applied the strictest criterion possible for cluster membership which states that, the descriptors for which their entire seven-digit codes were exactly the same could cluster together. The strong clusters obtained with their respective seven-digit codes are given in Table 4. So, for example, we see that the first descriptor **Ho**(**A**,*A*), the second one **Ho**(**A**,*E*), and the sixth one **Ho**(**A**,*Z*) were always included in the same cluster no matter what method was followed. They belonged to cluster number 1 for the average linkage, complete linkage, single linkage, Ward's method, Jarvis−Patrick, and variable Jarvis−Patrick, and in cluster number 14 for relocation method (it must be noted here that the cluster numberings in all different methods do not have any significant inherent meaning). Hence their codes are 1−1−1−1−1−1−14. However, if we consider, e.g., the descriptor **Wi**(**D**,^{4}*A*), having been assigned the code 1−1−1−1−20−1−14, belongs to cluster number 20 for the fourth CA method (Jarvis−Patrick), whereas for the other six methods it was in the same cluster with descriptors **Ho**(**A**,*A*), **Ho**(**A**,*E*), and **Ho**(**A**,*Z*). But according to the strict criterion employed, **Wi**(**D**^{4},*A*), could not cluster with these three, although it might look statistically quite probable that it might as well do so.

(ii) Second Step: Construction of Weak Clusters. As pointed out before, different CA methods have inherent biases and can generally behave differently. It is not possible to state that a single given method is better over another. The number of strong clusters obtained in step 1 (79) is rather high. Therefore, we were interested in exploring further groupings of descriptors, that is, larger clusters within this initially obtained structure. However, we tried to avoid building an artificial structure, by forcing further fusions of clusters that could be very unreliable. It is obvious that algorithms which would attempt to fuse strong clusters from step 1 would be order-dependent, thus producing groupings

Table 4. Compositions of Strong Clusters: Members of the Clusters that Contain Descriptors Which Invariably Belong to the Same Cluster in all Seven CA Methods Employed[a]

cluster	cardinality	code	descriptors
1	12	15−22−1−21−8−1−7	**HyWi(D,*A*)**, **HyWi(D,*E*)**, **HyWi(D,*AH*)**, **HyWi(D,*P*)**, **HyWi(D,*R*)**, **HyWi(D,*Z*)**, **Wi(D^2,*A*)**, **Wi(D^2,*E*)**, **Wi(D^2,*AH*)**, **Wi(D^2,*P*)**, **Wi(D^2,*R*)**, **Wi(D^2,*Z*)**
2	11	24−24−1−25−6−16−2	**Wi(A^3,*A*)**, **Wi(A^3,*E*)**, **Wi(A^3,*AH*)**, **Wi(A^3,*Z*)**, **Wi(A^4,*A*)**, **Wi(A^4,*E*)**, **Wi(A^4,*AH*)**, **Wi(A^4,*Z*)**, **^{2}Chi(RDS,*A*)**, **^{2}Chi(RDS,*E*)**, **^{2}Chi(RDS,*AH*)**
3	11	26−15−8−9−13−9−23	**MinSp(A,*A*)**, **MinSp(A,*E*)**, **MinSp(A,*AH*)**, **MinSp(A,*Z*)**, **MaxSp(A,*A*)**, **MaxSp(A,*E*)**, **MaxSp(A,*AH*)**, **MaxSp(A,*Z*)**, **^{3}Chi(RDS,*A*)$_p$**, **^{3}Chi(RDS,*E*)$_p$**, **^{3}Chi(RDS,*AH*)$_p$**
4	10	1−2−1−2−20−1−14	**Wi(D^3,*A*)**, **Wi(D^3,*E*)**, **Wi(D^3,*AH*)**, **Wi(D^3,*P*)**, **Wi(D^3,*R*)**, **Wi(D^3,*Z*)**, **Wi(D^4,*AH*)**, **Wi(D^4,*P*)**, **Wi(D^4,*R*)**, **Wi(D^5,*P*)**
5	9	16−25−1−26−9−1−25	**HyWi(RD,*A*)**, **HyWi(RD,*AH*)**, **HyWi(RD,*Z*)**, **Wi(RD^2,*A*)**, **Wi(RD^2,*AH*)**, **Wi(RD^2,*Z*)**, **Wi(RD^3,*AH*)**, **^{3}Chi(Val,*A*)$_p$**, **^{3}Chi(Val,*AH*)$_p$**
6	8	1−1−1−1−20−1−14	**Wi(D^4,*A*)**, **Wi(D^4,*E*)**, **Wi(D^4,*Z*)**, **Wi(D^5,*A*)**, **Wi(D^5,*E*)**, **Wi(D^5,*AH*)**, **Wi(D^5,*R*)**, **Wi(D^5,*Z*)**
7	8	13−20−1−20−10−1−12	MW **HyWi(RD,*E*)**, **Wi(RD,*E*)**, **^{0}Chi(Val,*E*)**, **^{0}Chi(Val,*AH*)**, **^{0}Chi(Val,*P*)**, **^{0}Chi(Val,*R*)**, **^{1}Chi(Val,*E*)**
8	8	13−20−1−19−10−1−12	**$^0\chi$**, **$^1\chi$**, **^{0}Chi(RDS,*E*)**, **^{0}Chi(RDS,*Z*)**, **^{0}Chi(Val,*Z*)**, **^{1}Chi(RDS,*Z*)**, **^{1}Chi(Val,*Z*)**, **^{2}Chi(Val,*Z*)**
9	8	17−28−1−27−10−1−28	**MaxSp(RD,*R*)**, **Wi(RD^2,*P*)**, **Wi(RD^3,*P*)**, **Wi(RD^4,*R*)**, **Wi(RD^5,*R*)**, **^{2}Chi(Val,*P*)**, **^{3}Chi(Val,*P*)$_p$**, **^{3}Chi(Val,*R*)$_p$**
10	6	13−21−1−20−10−1−11	**HyWi(RD,*P*)**, **HyWi(RD,*R*)**, **Wi(RD,*P*)**, **Wi(RD^2,*R*)**, **^{1}Chi(Val,*P*)**, **^{2}Chi(Val,*R*)**
11	6	13−21−1−20−10−1−18	**Wi(RD,*A*)**, **Wi(RD,*AH*)**, **Wi(RD,*Z*)**, **^{1}Chi(Val*A*)**, **^{1}Chi(Val,*AH*)**, **^{2}Chi(Val,*E*)**
12	6	15−19−1−22−15−1−21	**MaxSp(D,*A*)**, **MaxSp(D,*E*)**, **MaxSp(D,*AH*)**, **MaxSp(D,*P*)**, **MaxSp(D,*R*)**, **MaxSp(D,*Z*)**
13	6	18−10−1−14−9−1−27	**MaxSp(RD,*A*)**, **MaxSp(RD,*AH*)**, **MaxSp(RD,*Z*)**, **Wi(RD^5,*A*)**, **Wi(RD^5,*AH*)**, **Wi(RD^5,*Z*)**
14	6	27−16−17−15−24−19−4	**^{1}Chi(DS,*A*)**, **^{1}Chi(DS,*E*)**,**^{1}Chi(DS,*AH*)**, **^{1}Chi(DS,*P*)**, **^{1}Chi(DS,*R*)**, **^{1}Chi(DS,*Z*)**
15	5	12−18−1−18−17−1−9	**Wi(A,*A*)**, **Wi(A,*E*)**, **Wi(A,*AH*)**, **Wi(A,*R*)**, **Wi(A,*Z*)**
16	5	12−18−1−24−18−1−13	**Wi(A^2,*A*)**, **Wi(A^2,*E*)**, **Wi(A^2,*AH*)**, **Wi(A^2,*R*)**, **Wi(A^2,*Z*)**
17	5	12−19−1−18−28−1−9	**^{0}Chi(RDS,*P*)**, **^{1}Chi(RDS,*A*)**, **^{1}Chi(RDS,*E*)**, **^{1}Chi(RDS,*AH*)**, **^{1}Chi(RDS,*R*)**
18	5	15−19−1−22−14−1−7	**MinSp(D,*A*)**, **MinSp(D,*E*)**, **MinSp(D,*P*)**, **MinSp(D,*R*)**, **MinSp(D,*Z*)**
19	5	27−16−17−15−24−19−26	**^{2}Chi(DS,*E*)**, **^{2}Chi(DS,*R*)**, **^{2}Chi(DS,*Z*)**, **^{3}Chi(DS,*E*)$_p$**, **^{3}Chi(DS,*Z*)$_p$**
20	4	15−19−1−22−19−1−21	**Wi(D,*A*)**, **Wi(D,*E*)**, **Wi(D,*R*)**, **Wi(D,*Z*)**
21	4	17−28−1−27−10−1−6	**MaxSp(RD,*E*)**, **Wi(RD^4,*E*)**, **Wi(RD^5,*E*)**, **^{3}Chi(Val,*E*)$_p$**
22	4	27−17−17−16−24−19−5	**^{2}Chi(DS,*A*)**, **^{2}Chi(DS,*AH*)**, **Chi(DS,*A*)**, **^{3}Chi(DS,*AH*)$_p$**
23	3	1−1−1−1−1−1−14	**Ho(A,*A*)**, **Ho(A,*E*)**, **Ho(A,*Z*)**
24	3	8−9−18−13−16−12−27	**MinSp(RD,*A*)**, **MinSp(RD,*AH*)**, **MinSp(RD,*Z*)**
25	3	12−18−1−18−4−1−9	**HyWi(A,*A*)**, **HyWi(A,*E*)**, **HyWi(A,*Z*)**
26	3	13−20−1−19−10−1−9	**^{0}Chi(RDS,*A*)**, **^{0}Chi(RDS,*AH*)**, **^{0}Chi(RDS,*R*)**
27	3	17−26−1−28−10−1−28	**MaxSp(RD,*P*)**, **Wi(RD^4,*P*)** **Wi(RD^5,*P*)**
28	3	18−27−1−14−9−1−10	**Wi(RD^4,*A*)**, **Wi(RD^4,*AH*)**, **Wi(RD^4,*Z*)**
29	3	19−27−5−11−26−21−10	**^{3}Chi(Val,*E*)**c, **^{3}Chi(Val,*P*)$_c$**, **^{3}Chi(Val,*R*)$_c$**
30	3	21−11−20−12−23−18−22	**^{0}Chi(DS,*E*)**, **^{0}Chi(DS,*P*)**, **^{0}Chi(DS,*R*)**
31	3	25−14−10−10−12−26−1	**^{3}Chi(RDS,*E*)$_c$**, **^{3}Chi(RDS,*R*)$_c$**, **^{3}Chi(RDS,*Z*)$_c$**
32	3	27−17−17−16−24−19−17	**^{3}Chi(DS,*E*)$_c$**, **^{3}Chi(DS,*R*)$_c$**, **^{3}Chi(DS,*Z*)$_c$**
33	2	12−18−1−18−7−1−9	**HyWi(A,*R*)**, **Wi(A,*P*)**
34	2	13−20−1−20−10−1−11	**Wi(RD,*R*)**, **^{1}Chi(Val,*R*)**
35	2	16−25−1−26−9−1−10	**Wi(RD^3,*A*)**, **Wi(RD^3,*Z*)**
36	2	16−27−1−26−9−1−10	**^{2}Chi(Val,*A*)**, **^{2}Chi(Val,*AH*)**
37	2	18−10−4−13−25−1−27	**^{3}Chi(Val,*A*)$_c$**, **^{3}Chi(Val,*AH*)$_c$**
38	2	21−11−19−12−23−18−22	**^{0}Chi(DS,*A*)**, **^{0}Chi(DS,*AH*)**
39	2	23−23−1−23−6−5−24	**HyWi(A,*P*)**, **Wi(A^2,*P*)**
40	2	24−24−1−23−6−16−2	**Wi(A^3,*R*)**, **Wi(A^4,*R*)**
41	2	26−15−12−7−6−9−23	**MaxSp(A,*R*)**, **^{3}Chi(RDS,*R*)$_p$**
42	2	26−15−9−10−12−9−16	**^{3}Chi(RDS,*A*)$_c$**, **^{3}Chi(RDS,*AH*)$_c$**
43	2	28−17−17−17−24−19−15	**^{3}Chi(DS,*A*)$_c$**, **^{3}Chi(DS,*AH*)$_c$**
44	1	2−3−25−3−2−2−8	**Ho(A,*AH*)**
45	1	3−3−26−3−2−4−8	**Ho(A,*R*)**
46	1	4−8−28−4−3−3−8	**Ho(A,*P*)**
47	1	5−4−11−7−6−10−8	**MinSp(A,*P*)**
48	1	5−4−11−7−6−11−8	**MaxSp(A,*P*)**
49	1	5−4−13−7−6−11−24	**Wi(A^4,*P*)**
50	1	6−4−16−8−6−24−8	**^{3}Chi(RDS,*P*)$_p$**
51	1	7−5−14−8−6−23−8	**^{2}Chi(RDS,*P*)**
52	1	7−5−15−8−6−25−8	**^{3}Chi(RDS,*P*)$_c$**
53	1	9−7−23−6−16−14−3	**MinSp(RD,*P*)**
54	1	10−7−24−6−16−15−3	**MinSp(RD,*R*)**
55	1	11−6−27−5−16−13−3	**MinSp(RD,*E*)**
56	1	12−18−1−18−5−1−9	**HyWi(A,*AH*)**
57	1	12−20−1−19−27−1−12	**^{3}Chi(Val,*Z*)$_p$**
58	1	13−21−1−20−10−1−6	**Wi(RD^2,*E*)**
59	1	13−20−1−19−29−1−12	**^{2}Chi(RDS,*Z*)**
60	1	13−21−1−27−10−1−6	**Wi(RD^3,*E*)**
61	1	13−20−1−20−10−1−18	**^{0}Chi(Val,*A*)**
62	1	14−28−3−24−11−6−20	**$^2\chi$**
63	1	14−28−3−24−11−7−20	**$^3\chi_p$**
64	1	15−19−1−22−14−1−21	**MinSp(D,*AH*)**
65	1	15−19−1−22−21−1−21	**Wi(D,*AH*)**

Table 4. (Continued)

cluster	cardinality	code	descriptors
66	1	15−19−1−22−22−1−21	**Wi**(**D**,*P*)
67	1	17−28−1−27−10−1−11	**Wi**(**RD**^{3},*R*)
68	1	20−13−6−11−26−22−20	^{3}**Chi**(**Val**,*Z*)$_c$
69	I	21−11−21−12−23−20−22	^{0}**Chi**(**DS**,*Z*)
70	1	22−12−22−12−23−28−22	^{3}**Chi**(**RDS**,*Z*)$_p$
71	1	23−23−2−23−6−5−24	^{1}**Chi**(**RDS**,*P*)
72	1	24−24−1−23−6−17−24	**Wi**(**A**^{3},*P*)
73	1	24−24−1−23−6−27−24	^{2}**Chi**(**RDS**,*R*)
74	1	25−13−7−11−12−8−1	$^3\chi_c$
75	1	26−15−12−7−6−11−23	**MinSp**(**A**,*R*)
76	1	27−16−17−15−24−19−19	^{3}**Chi**(**DS**,*R*)$_p$
77	1	27−16−17−16−24−19−19	^{2}**Chi**(**DS**,*P*)
78	I	27−17−17−16−24−19−19	^{3}**Chi**(**DS**,*P*)$_p$
79	1	27−17−17−16−24−19−15	^{3}**Chi**(**DS**,*P*)$_c$

[a] The seven-digit codes are also cited. The ranking of clusters depends only on their sizes (cardinality).

Table 5. Behavior of the Different CA Methods, When Applying Progressively More Relaxed Rules for Cluster Membership of the Descriptors[a]

ca method	6 out of 7	5 out of 7	4 out of 7
average linkage	0	6	14
complete linkage	4	45	59
single linkage	4	6	13
Ward's	11	32	63
Jarvis−Patrick	35	50	70
variable Jarvis−Patrick	5	9	13
K-means (relocation)	40	74	78

[a] Clustered together by differing in a single method (second column), two methods (third column) and three methods (fourth column).

that may not be, in principle, unique. That is different ordering of the objects (descriptors), or a totally by chance treatment of them might as well lead to different groupings. For this reason we followed an alternative approach. By relaxing the cluster-membership criterion, we allowed groupings obeying a 6 out of 7 rule; that is, the clustered descriptors had the same cluster appurtenance not necessarily in all seven methods, but in at least six, any of them. For example, we consider the cluster membership codes for **MinSp**(**A**,*P*) and **MaxSp**(**A**,*P*) which belong respectively to strong clusters 47 and 48 (Table 4). Their respective codes are 5−4−11−7−6−10−8 and 5−4−11−7−6−11−8, thereby differing in only one method, the variable Jarvis−Patrick. Application of this less strict criterion, which allowed descriptors such as these to be grouped together, resulted in a reduction of the number of descriptor clusters to 51. It was interesting, though, that the behavior of the different CA methods was not the same. In the above-cited example the two descriptors had variability in one position of their codes, an outcome of the variable Jarvis−Patrick method. Thus, we examined all such variabilities of cluster membership of individual descriptors in all seven methods to determine which methods made most of the difference. For average linkage clustering, there was not a single case in which any descriptors differed in only this method. On the contrary, this occurred in 40 cases for the relocation method. So, there were 40 cases of descriptors which were in the same clusters in all CA methods but for the relocation method. The results for the seven different methods are given in the second column of Table 5. By further relaxing the condition for cluster inclusion of descriptors to 5 out of 7 methods in common, we obtained 40 groups, and the membership differences as they were attributed to the different methods are given in the third column of Table 5. Last, the same procedure was repeated by applying a 4 out of 7 rule, which produced 30 groups with the corresponding differences in cluster membership due to the various methods, as shown in the fourth column of Table 5. These results clearly showed that the most consistent behavior had the three methods: average linkage, single linkage, and variable Jarvis−Patrick. Therefore, by considering only these three methods, we applied as final criterion of cluster membership the requirement that any descriptors in order to cluster together to have 3 out of 3 in common in their original seven-digit codes in exactly these three methods. The clusters thus obtained, referred to as *weak* clusters, are presented in Table 6.

DISCUSSION

The weak clusters of descriptors presented in Table 6 show clearly that there are certain groups of descriptors which tend to be always together. In particular, if we depend on the weighting scheme across the different operators, irrespective of the matrix employed for the derivation of the descriptors, there is a strong tendency of descriptors derived from the weighting schemes *A* and AH to cluster together. In 37 cases (31 in strong clusters and 6 in weak-only clusters) we have the descriptors of the same operator derived by means of weighting schemes *A* and AH belonging to the same cluster. The second strongest tendency to cluster together is found for weighting schemes *E* and *R* in which case 30 times (16 for strong clusters and 14 for weak-only clusters) descriptors derived from the same operator and these two weighting schemes cluster together. The third strongest such case of interrelationship among the weighting schemes is the pair *A* and *Z* of the weighting schemes, in which case they cluster together in 29 cases (24 strong and 5 weak-only). On the other hand, the weakest pairwise relationships are shown by the pairs of weighting schemes AH−*P* and *A*−*P* for which descriptors derived using the same operator and these two weighting schemes belong to the same cluster in only 14 cases (7 strong clusters and 7 weak-only ones for the first pair and 6 strong and 8 weak clusters for the second pair). Immediately follows the pair *Z* and *P* in only 15 cases (7 strong clusters and 8 weak-only ones). From these results (the complete list in Table 7) it is evident that, for the same operator, the weighting scheme *P* tends to give less redundant

Table 6. Composition of Weak Clusters[a]

cluster	cardinality	descriptors
1	37	**MolWeight**, $^0\chi$, $^1\chi$, **HyWi(RD**,*E*), **HyWi(RD**,*P*), **HyWi(RD**,*R*), **Wi(RD,***A*), **Wi(RD**,*E*), **Wi(RD**,*AH*), **Wi(RD**,*P*), **Wi(RD**,*R*), **Wi(RD**^{2},*E*), **Wi(RD**,*Z*), **Wi(RD**^{2},*R*), **Wi(RD**^{3},*E*), ^{0}**Chi(RDS**,*A*), ^{0}**Chi(RDS**,*AH*), ^{0}**Chi(RDS**,*R*), ^{0}**Chi(RDS**,*E*), ^{0}**Chi(RDS**,*Z*), ^{1}**Chi(RDS**,*Z*), ^{0}**Chi(Val**,*A*), ^{0}**Chi(Val**,*E*), ^{0}**Chi(Val**,*AH*), ^{0}**Chi(Val**,*P*), ^{0}**Chi(Val**,*R*), ^{0}**Chi(Val**,*Z*), ^{1}**Chi(Val**,*A*), **Chi(Val**,*E*), ^{1}**Chi(Val**,*AH*), ^{1}**Chi(Val**,*P*), ^{1}**Chi(Val**,*R*), ^{1}**Chi(Val**,*Z*), ^{2}**Chi(Val**,*E*), ^{2}**Chi(Val**,*R*), ^{2}**Chi(Val**,*Z*), ^{2}**Chi(RDS**,*Z*)
2	30	**HyWi(D**,*A*), **HyWi(D**,*E*), **HyWi(D**,*AH*), **HyWi(D**,*P*), **HyWi(D**,*R*), **HyWi(D**,*Z*), **Wi(D**,*A*), **Wi(D**,*E*), **Wi(D**,*AH*), **Wi(D**,*P*), **Wi(D**,*R*), **Wi(D**,*Z*), **Wi(D**^{2},*A*), **Wi(D**^{2},*E*), **Wi(D**^{2},*AH*), **Wi(D**^{2},*P*), **Wi(D**^{2},*R*), **Wi(D**^{2},*Z*), **MinSp(D**,*A*), **MinSp(D**,*E*), **MinSp(D**,*P*), **MinSp(D**,*R*), **MinSp(D**,*Z*), **MinSp(D**,*AH*), **MaxSp(D**,*A*), **MaxSp(D**,*E*), **MaxSp(D**,*AH*), **MaxSp(D**,*P*), **MaxSp(D**,*R*), **MaxSp(D**,*Z*)
3	22	**HyWi(A**,*A*), **HyWi(A**,*E*), **HyWi(A**,*AH*), **HyWi(A**,*R*), **HyWi(A**,*Z*), **Wi(A**,*A*), **Wi(A**,*E*), **Wi(A**,*AH*), **Wi(A**,*R*), **Wi(A**,*Z*), **Wi(A**,*P*), **Wi(A**^{2},*A*), **Wi(A**^{2},*E*), **Wi(A**^{2},*AH*), **Wi(A**^{2},*R*), **Wi(A**^{2},*Z*), ^{0}**Chi(RDS**,*P*), ^{1}**Chi(RDS**,*A*), ^{1}**Chi(RDS**,*E*), **Chi(RDS**,*AH*), ^{1}**Chi(RDS**,*R*), ^{3}**Chi(Val**,*Z*)$_p$
4	22	^{2}**Chi(DS**,*A*), ^{1}**Chi(DS**,*E*), **Chi(DS**,*AH*), **Chi(DS**,*P*), ^{1}**Chi(DS**,*R*), ^{1}**Chi(DS**,*Z*), ^{1}**Chi(DS**,*A*), ^{2}**Chi(DS**,*E*), ^{2}**Chi(DS**,*AH*), ^{2}**Chi(DS**,*P*), ^{2}**Chi(DS**,*R*), ^{2}**Chi(DS**,*Z*), ^{1}**Chi(DS**,*A*)$_p$, ^{3}**Chi(DS**,*E*)$_p$, ^{3}**Chi(DS**,*AH*)$_p$, ^{3}**Chi(DS**,*P*)$_p$, ^{3}**Chi(DS**,*R*)$_p$, ^{3}**Chi(DS**,*Z*)$_p$, ^{3}**Chi(DS**,*E*)$_c$, ^{3}**Chi(DS**,*P*)$_c$ ^{3}**Chi(DS**,*R*), ^{3}**Chi(DS**,*Z*)$_c$
5	21	**Ho(A**,*A*), **Ho(A**,*E*), **Ho(A**,*Z*), **Wi(D**^{3},*A*), **Wi(D**^{4},*A*), **Wi(D**^{5},*A*), **Wi(D**^{3},*E*), **Wi(D**^{4},*E*), **Wi(D**^{5},*E*), **Wi(D**^{3},*AH*), **Wi(D**^{4},*AH*), **Wi(D**^{5},*AH*), **Wi(D**^{3},*P*), **Wi(D**^{4},*P*), **Wi(D**^{5},*P*), **Wi(D**^{3},*R*), **Wi(D**^{4},*R*), **Wi(D**^{5},*R*), **Wi(D**^{3},*Z*), **Wi(D**^{4},*Z*), **Wi(D**^{5},*Z*)
6	16	**MaxSp(RD**,*E*), **MaxSp(RD**,*P*), **MaxSp(RD**,*R*), **Wi(RD**^{2},*P*), **Wi(RD**^{3},*R*), **Wi(RD**^{4},*E*), **Wi(RD**^{4},*P*), **Wi(RD**4,*R*), **Wi(RD**^{5},*E*), **Wi(RD**^{5},*P*), **Wi(RD**^{5},*R*), ^{2}**Chi(Val**,*P*), ^{3}**Chi(Val**,*E*)$_p$, ^{3}**Chi(Val**,*P*)$_p$, ^{3}**Chi(Val**,*R*)$_p$
7	13	**HyWi(RD**,*A*), **HyWi(RD**,*AH*), **HyWi(RD**,*Z*), **Wi(RD**^{2},*A*), **Wi(RD**^{2},*AH*), **Wi(RD**^{2},*Z*), **Wi(RD**^{3},*A*), **Wi(RD**^{3},*AH*), **Wi(RD**^{3},*Z*), ^{2}**Chi(Val**,*A*), ^{2}**Chi(Val**,*AH*), ^{3}**Chi(Val**,*A*)$_p$, ^{3}**Chi(Val**,*AH*)$_p$
8	13	**Wi(A**^{3},*A*), **Wi(A**^{3},*E*), **Wi(A**^{3},*AH*), **Wi(A**^{3},*R*), **Wi(A**^{3},*Z*), **Wi(A**^{4},*A*), **Wi(A**^{4},*E*), **Wi(A**^{4},*AH*), **Wi(A**^{4},*R*), **Wi(A**^{4},*Z*), ^{2}**Chi(RDS**,*A*), ^{2}**Chi(RDS**,*E*), ^{2}**Chi(RDS**,*AH*)
9	11	**MinSp(A**,*A*), **MinSp(A**,*E*), **MinSp(A**,*AH*), **MinSp(A**,*Z*), **MaxSp(A**,*A*), **MaxSp(A**,*E*), **MaxSp(A**,*AH*), **MaxSp(A**,*Z*), ^{3}**Chi(RDS**,*A*)$_p$, ^{3}**Chi(RDS**,*E*)$_p$, ^{3}**Chi(RDS**,*AH*)$_p$
10	9	**MaxSp(RD**,*A*), **MaxSp(RD**,*AH*), **MaxSp(RD**,*Z*), **Wi(RD**^{4},*A*), **Wi(RD**^{4},*AH*), **Wi(RD**^{4},*Z*), **Wi(RD**^{5},*A*), **Wi(RD**^{5},*AH*), **Wi(RD**^{5},*Z*)
11	3	**MinSp(RD**,*A*), **MinSp(RD**,*AH*), **MinSp(RD**,*Z*)
12	3	^{3}**Chi(Val**,*E*)$_c$, ^{3}**Chi(Val**,*P*)$_c$, ^{3}**Chi(Val**,*R*)$_c$
13	3	^{0}**Chi(DS**,*E*), ^{0}**Chi(DS**,*P*), ^{0}**Chi(DS**,*R*)
14	3	^{3}**Chi(RDS**,*E*)$_c$, ^{3}**Chi(RDS**,*R*)$_c$, ^{3}**Chi(RDS**,*Z*)$_c$
15	2	^{3}**Chi(Val**,*A*)$_c$, ^{3}**Chi(Val**,*AH*)$_c$
16	2	^{0}**Chi(DS**,*A*), ^{0}**Chi(DS**,*AH*)
17	2	**HyWi(A**,*P*), **Wi(A**^{2},*P*)
18	2	**MaxSp(A**,*R*), ^{3}**Chi(RDS**,*R*)$_p$
19	2	^{3}**Chi(RDS**,*A*)$_c$, ^{3}**Chi(RDS**,*AH*)$_c$
20	2	^{3}**Chi(DS**,*A*)$_c$, ^{3}**Chi(DS**,*AH*)$_c$
21	1	**Ho(A**,*AH*)
22	1	**Ho(A**,*P*)
23	1	**Ho(A**,*R*)
24	1	**MinSp(A**,*P*)
25	1	**MaxSp(A**,*P*)
26	1	**Wi(A**^{4},*P*)
27	1	^{3}**Chi(RDS**,*P*)$_p$
28	1	^{2}**Chi(RDS**,*P*)
29	1	^{3}**Chi(RDS**,*P*)$_c$
30	1	**MinSp(RD**,*P*)
31	1	**MinSp(RD**,*R*)
32	1	**MinSp(RD**,*E*)
33	1	$^2\chi$
34	1	$^3\chi_p$
35	1	^{3}**Chi(Val**,*Z*)$_c$
36	1	^{0}**Chi(DS**,*Z*)
37	1	^{3}**Chi(RDS**,*Z*)$_p$
38	1	^{1}**Chi(RDS**,*P*)
39	1	**Wi(A**^{3},*P*)
40	1	^{2}**Chi(RDS**,*R*)
41	1	$3\chi_c$
42	1	**MinSp(A**,*R*)

[a] The members of these clusters contain descriptors which belong to the same cluster in the 3 most consistent out of totally 7 CA method employed.

descriptors with respect to the other weighting schemes. These findings are corroborated by the fact that descriptors derived with the weighting scheme *P* appear in 15 cases in weak clusters and an additional 11 cases in strong clusters in which no other descriptor derived from the same operator does. That is, the descriptors derived with *P* seem to be the less correlated with the descriptors derived from the same operator but using another weighting scheme. These findings are shown in Table 8, in which we see that, for weak clusters only, in only 6 cases descriptors derived from weighting schemes *Z* and *R* show the same behavior, in 2 cases for weighting scheme *E* and, finally, in 1, and none for AH and *A* respectively. The conclusions that we can draw, by examining the groupings of descriptors in Table 6, as far as the operators employed for the calculation of descriptors are concerned, are the following:

The descriptors derived with the **HyWi** operator generally cluster together with those derived from the **Wi** operator of orders 1 and 2, for matrices **A**, **RD**, and **D**. In addition, for the case of matrix **RD**, the **HyWi**-based descriptors also

Table 7. Intercorrelations of Weighting Schemes Across the Various Operators by Pairwise Comparisons of the Cases in Which Descriptors Derived from the Same Operator and the Weighting Schemes in the Pair Belong to the Same Cluster

pair of weighting schemes	strong clusters	weak-only clusters	pair of weighting schemes	strong clusters	weak-only clusters
A−E	20	5	E−Z	22	3
A−AH	31	6	AH−P	7	7
A−P	6	8	AH−R	12	8
A−R	12	8	AH−Z	18	10
A−Z	24	5	P−R	13	10
E−AH	16	8	P−Z	7	8
E−P	11	12	R−Z	13	9
E−R	16	14			

Table 8. Times Descriptors Derived from a Given Weighting Scheme and Any Operator *Not* Belong to the Same Cluster with the Rest of the Descriptors Derived from the Same Operator, but with Different Weighting Schemes[a]

weighting scheme	strong clusters only	weak-only clusters	weighting scheme	strong clusters only	weak-only clusters
A	1	0	P	11	15
E	9	2	R	12	6
AH	4	1	Z	3	6

[a] Second column refers to strong clusters only; third column includes the *additional* cases in which weak clusters are considered only.

cluster together with the third order **Wi** operators. Moreover the descriptors derived with the **HyWi** operator and the **A** matrix tend to have more in common with the **Chi** indices of first order and the **RDS** matrix, whereas for the **HyWi** operator with the matrix **RD** the same applies to the case of the **Val** matrix, for orders 1, 2, and 4 of these **Chi** indices, as well as for the 0th order **Chi** index derived by using the **RDS** matrix. However, it is also quite clear that the descriptors derived with the **HyWi** operator are close to those descriptors derived with the **Wi** operator and the same matrix (weak clusters 1, 2, 3, 7, and 17).

For the descriptors **MinSp**, **MaxSp** and the different distance matrices and weighting schemes, in some cases there is not clear separation between **MinSp**- and the **MaxSp**-derived descriptors (as is the case in weak clusters 2 and 9). In other cases in which there is such a separation according to the criterion of formation of weak clusters, e.g., **MinSp**(**A**,*P*), and **MaxSp**(**A**,*P*), these two still tend to cluster together in 6 out of the 7 CA methods. Very good separation exists for those descriptors derived with the **RD** matrix (weak clusters 11 and 30−32 as well as 6 and 10). Also, the descriptors derived from the **MaxSp** operator tend to be much closer to **Wi** indices derived with the same weighting schemes and matrices, e.g., weak clusters 2, 6, and 10.

The descriptors derived with the **Ho** operator are much closer to the **Wi** operators of higher orders (3−5) although derived from different matrices. This is more profound for the weighting schemes *A*, *E*, and *Z* in weak cluster 5. In the very same weak cluster we can also see the close relationships between the descriptors of **Wi** operators of higher orders.

The **Chi** indices of lower orders (0,1) derived from the same matrix are quite close to each other (weak clusters 1, 3, 13, 16, 36, 37, and 4). For the third order **Chi** indices, those of **p** type cluster together and separately from those of **c** type. This is true no matter what matrix is employed, which can be seen in weak clusters 27 and 29, 3 and 35, 7 and 15, 6 and 12, and 4 and 20.

It is also interesting to note the behavior of "outlier" descriptors. Looking at the behavior of the descriptors, across all 7 CA methods, we can see that from the 79 clusters initially obtained there are 34 which contain only 1 descriptor. If we were to consider this behavior horizontally, that is, by ignoring the fact that we deal with 7 different CA methods which we could even consider as a single CA method operating in 7 stages, we might as well refer to these clusters technically as singletons, since the objects they contain tend to be rather different when compared to the other ones. However, these clusters are not necessarily always vertical singletons, that is, singletons within each single method. In fact, the results in Tables 3 and 4 show that there is not a single descriptor at this clustering level which is considered singleton simultaneously in all 7 methods employed. E.g., descriptor **Ho**(**A**,*P*), which is, by far, the least correlated and, by definition, the most distinct object in the entire collection, does indeed form a singleton cluster in all but one method (relocation). It goes that by considering the 6-out-of-7-common rule it is the only descriptor which can be seen as "pure" singleton. According to the 5-out-of-7-common rule, there is an additional descriptor which can be considered singleton, the **MinSp**(**RD**,*E*), and by applying the 4-out-of-7-common rule there also appears as singleton the descriptor ^{3}**Chi**(**RDS**,*Z*)$_{p}$. In the case of the 42 weak clusters obtained according to the criterion of the three mutually consistent methods (average linkage, single linkage, and variable Jarvis−Patrick) we can see in Table 6 that there are 22 clusters which contain only one (outlier) descriptor. However, if we look vertically at each of these three methods, there are only 9 descriptors which form singleton clusters in all of these methods. They are **Ho**(**A**,*AH*), **Ho**(**A**,*P*), **Ho**(**A**,*R*), **MinSp**(**RD**,*E*), **MinSp**(**RD**,*P*), **MinSp**(**RD**,*R*), ^{3}**Chi**(**Val**,*Z*)$_{c}$, ^{3}**Chi**(**RDS**,*P*)$_{p}$, and ^{3}**Chi**(**RDS**,*Z*)$_{p}$. Moreover, if we examine the doubleton weak clusters, i.e., clusters which contain only two members, which are ^{3}**Chi**(**Val**,*A*)$_{c}$ and ^{3}**Chi**(**Val**,*AH*)$_{p}$, ^{0}**Chi**(**DS**,*A*) and ^{0}**Chi**(**DS**,*AH*), **HyWi**(**A**,*P*) and **Wi**(**A**^{2},*P*), **MaxSp**(**A**,*R*) and ^{3}**Chi**(**RDS**,*R*)$_{p}$, ^{3}**Chi**(**RDS**,*A*)$_{c}$ and ^{3}**Chi**(**RDS**,*AH*)$_{c}$, ^{3}**Chi**(**DS**,*A*)$_{c}$, and ^{3}**Chi**(**DS**,*AH*)$_{c}$, we see that these pairs contain in all but one case descriptors derived from the same matrix, from identical or closely related operators and from the same or very correlated weighting schemes (as is the case, for example, of *A* and AH, as seen in Table 7).

Also in the case of the singleton clusters, the descriptors are quite well related with those in the doubleton clusters, indicating that there are certain operators which, in fact, have a tendency to sample distant areas of the multivariate space. These are **Ho, MinSp,** and the zero and third order **Chi** descriptors. These findings are also confirmed by the "tripleton" (three-member) clusters, that is, the weak clusters 11−14.

CONCLUSIONS

The large number of available molecular descriptors, and the fact that many of them are highly correlated, confounds the development of both predictive models and the assessment of molecular similarity/dissimilarity. Therefore, it is

desirable to employ smaller, more manageable, sets of relatively independent (as orthogonal as possible) variables, thereby ensuring that these descriptors contain nonredundant information.

In comparison with previous studies on clustering and similarity of molecular graph descriptors, in this article we investigate a much larger number of TIs computed with several weighting schemes, computed for a database of 2000 compounds tested in the AIDS program of NCI.

Instead of using the more traditional approaches of FA and PCA, the present work aimed at dividing a set of 240 graph-theoretical descriptors into nonoverlapping clusters, with the property that descriptors within a cluster are *related* to each other and are, at the same time, *unrelated* to those found in other clusters. CA methods are interested primarily in identifying groups of objects in whatever dimension space the objects are in, whereas FA methods are interested primarily in the dimension of the space the objects are in. There may be only a, e.g., 2D space as determined by FA, but there may be 1, 2, 3, or more meaningful clusters identified by a clustering method. Since the effect of FA is to transform the data so that any modes present are merged, resulting in variables that are normally distributed, FA may blur the relationship between groups of variables because it assumes that the factor scores are normally distributed. PCA, although it favors maintaining the representation of widely separated groups in a reduced space, it also minimizes, and thus may blur, the distances between clusters or groups that are not widely separated.

In our approach we can consider that the formed clusters constitute new classes of variables, conceptually much like the factors and PCs derived from FA and PCA. Then a single descriptor from each cluster may be selected as adequately representing all of the other variables included in its cluster. Descriptors from different clusters can be used either for optimal design of diverse libraries of chemical compounds or for the formulation of QSAR/QSPR models.

The results of this cluster analysis were in very good agreement with those obtained in an earlier work in which we addressed the issue of selecting noncorrelated (quasi-orthogonal) descriptors from the same set of graph-theoretical parameters by applying a heuristic algorithm.[35] This fact mutually confirmed the validity of both of these different strategies for retrieving individual uncorrelated descriptors from a much larger pool of computed descriptors.

In addition, intercorrelations between different operators, weighting schemes, and molecular matrices used for the calculation of these descriptors were derived on the basis of the formed clusters.

LIST OF ACRONYMS AND ABBREVIATIONS

A: adjacency matrix
A: atomic mass (weighting scheme)
AH: hydrogen-augmented atomic mass (weighting scheme)
CA: cluster analysis
CE: clustering entropy
Chi: chi indices
D: distance matrix
DEC: determination coefficient
E: atomic electronegativity (weighting scheme)
ENC: effective number of clusters
FA: factor analytical
HC: hierarchical clustering
Ho: Hosoya operator
HyWi: hyper-Wiener operator
M: molecular matrix
MaxSp: maximal spectrum eigenvalue operator
MinSp: minimal spectrum eigenvalue operator
MW: molecular weight
NHC: nonhierarchical clustering
P: atomic polarizability (weighting scheme)
PCA: principal component analysis
QSAR: quantitative structure−activity relationships
QSPR: quantitative structure−property relationships
R: atomic radius (weighting scheme)
RD: reciprocal distance matrix
RDS: reciprocal distance sum
TI: topological indices
Val: valency
VSD: vertex invariant
Wi: Wiener operator
Z: atomic number (weighting scheme)

REFERENCES AND NOTES

(1) Johnson, M. A., Maggiora, G. M., Eds. *Concepts and Applications of Molecular Similarity*; Wiley-Interscience: New York, 1990.
(2) Mason, J. S.; McLay, I. M.; Lewis, R. A. Applications of Computer-Aided Drug Design Techniques to Lead Generation. In *New Perspectives in Drug Design*; Dean, P. M., Jolles, G., Newton, C. G., Eds.; Academic Press: London, 1994; Chapter 12, pp 225−253.
(3) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sc*i. **1996**, *36*, 572−584.
(4) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.
(5) Brown, R. D. Descriptors for Diversity Analysis. *Perspect. Drug Des.* **1997**, *7/8*, 31−49.
(6) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand−Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.
(7) Matter, H. Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors *J. Med. Chem.* **1997**, *40*, 1219−1229.
(8) Martin, Y. C.; Bures, M. G.; Brown, R. D. Validated Descriptors for Diversity Measurements and Optimization. *Pharm. Pharmacol. Commun*. **1998**, *4*, 147−152.
(9) Pötter, T.; Matter, H. Random or Rational Design? Evaluation of Diverse Compound Subsets from Chemical Structure Databases. *J. Med. Chem*. **1998**, *41*, 478−488.
(10) Bayada, D. M.; Hamersma, H.; van Geerestein, V. J. Molecular Diversity and Representativity in Chemical Databases. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 1−10.
(11) Pepperrell, C. A.; Willett, P. Techniques for the Calculation of Three-Dimensional Structural Similarity Using Inter-Atomic Distances. *J. Comput.-Aided Mol. Des.* **1991**, *533*, 455−474.
(12) Brown, R. D.; Bures, M. G.; Martin, Y. C. A Comparison of Some Commercially Available Structural Descriptors and Clustering Algorithms. In *Proceedings of the First Electronic Computational Chemistry Conference*; Bachrach, S. M., Boyd, D. B., Gray, S. K., Hase, W., Rzepa, H. S., Eds.; ARInternet: Landover, MD, 1995.
(13) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical Similarity Using Geometric Atom Pair Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128−136.
(14) Briem H.; Kuntz I. D. Molecular Similarity Based on DOCK-Generated Fingerprints. *J. Med. Chem*. **1996**, *39*, 3401−3408.
(15) Van Geerestein, V. J.; Hamersma, H.; Van Helden, S. P. Exploiting Molecular Diversity: Pharmacophore Searching and Compound Clustering. Computer-Assisted Lead Finding and Optimization. In *Current Tools for Medicinal Chemistry*; Waterbeemd, H., Van de Testa, B., Folkers, G., Eds.; Wiley-VCH: Weinheim, Germany, 1997; pp 159−178.
(16) Nilakantan, R.; Bauman, N.; Haraki, K. S. Database Diversity Assessment: New Ideas, Concepts and Tools. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 447−452.
(17) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(18) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press: Letchworth, U.K., 1986.
(19) Ivanciuc, O.; Balaban, A. T. The Graph Description of Chemical Structures. In *Topological Indices and Related Descriptors in QSAR*

and QSPR; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: Amsterdam, The Netherlands, 1999; pp 59−167.
(20) Ivanciuc, O.; Ivanciuc, T.; Balaban, A. T. Vertex- and Edge-Weighted Molecular Graphs and Derived Structural Descriptors. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: Amsterdam, The Netherlands, 1999; pp 169−220.
(21) Basak, S. C.; Balaban, A. T.; Grunwald, G. D.; Gute, B. D. Topological Indices: Their Nature and Mutual Relatedness. *J. Chem. Inf. Comput. Sci.*, in press.
(22) Willett, P. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(23) Randić, M. Orthogonal Molecular Descriptors. *New J. Chem.* **1991**, *15*, 517−525.
(24) Lewis, R. A.; Mason, J. S.; McLay, I. M. Similarity Measures for Rational Set Selection and Analysis of Combinatorial Libraries: The Diverse Property-Derived (DPD) Approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 599−614.
(25) Stanton, D. T. Evaluation and Use of BCUT Descriptors in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 11−20.
(26) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. Determining Structural Similarity of Chemicals Using Graph-Theoretic Indices. *Discrete Appl. Math.* **1988**, *19*, 17−44.
(27) Malinkwoski, E. R.; Howery, D. G. *Factor Analysis in Chemistry*; John Wiley & Sons: New York, 1980.
(28) Motoc, I.; Balaban, A. T. Topological Indices: Intercorrelations, Physical Meaning, Correlational Ability. *Rev. Roum. Chim.* **1981**, *26*, 593−600.
(29) Motoc, I.; Balaban, A. T.; Mekenyan, O.; Bonchev, D. Topological Indices: Inter-Relations and Composition. *MATCH (Commun. Math. Chem.)* **1982**, *13*, 369−404.
(30) Kovacevic, K.; Plavšic, D.; Trinajstić, N.; Horvath, D. On the Intercorrelation of Topological Indices. In *MATH/CHEM/COMP 1988, Proceedings of an International Course and Conference on the Interfaces Between Mathematics, Chemistry and Computer Sciences*, Dubrovnik, Yugoslavia, June 20−25, 1988; Graovac, A., Ed.; Studies in Physical and Theoretical Chemistry, Vol. 63; Elsevier: Amsterdam, 1989; pp 213−224.
(31) Horvath, D.; Graovac, A.; Plavšic, D.; Trinajstić, N.; Strunje, M. On the Intercorrelation of Topological Indices in Benzenoid Hydrocarbons. *Int. J. Quantum Chem.: Quantum Chem. Symp.* **1992**, *26*, 401−408.
(32) Ivanciuc, O.; Ivanciuc, T.; Diudea, M. V. Molecular Graph Matrices and Derived Structural Descriptors. *SAR QSAR Environ. Res.* **1997**, *7*, 63−87.
(33) Ivanciuc, O.; Diudea, M. V.; Khadikar, P. V. New Topological Matrices and Their Polynomials. *Ind. J. Chem.* **1998**, *37A*, 574−585.
(34) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R.; Veith, G. D. Topological Indices: Their Nature, Mutual Relatedness, and Applications. *Math. Model.* **1987**, *8*, 300−305.
(35) Ivanciuc, O.; Taraviras, S. L.; Cabrol-Bass, D. Quasi-Orthogonal Basis Sets of Molecular Graph Descriptors as a Chemical Diversity Measure. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 126−134.
(36) Cunningham, K. M.; Ogilvie, J. C. Evaluation of Hierarchical Grouping Techniques. *Comput. J.* **1972**, *15*, 209−213.
(37) Sneath, P. H. A.; Sokal, R. R. *Numerical Taxonomy*; W. H. Freeman: San Francisco, 1973.
(38) Willett, P. *Similarity and Clustering in Chemical Information Systems*; John Wiley and Sons: New York, 1987.
(39) Downs, G. M.; Willett, P. The Use of Similarity and Clustering Techniques for the Prediction of Molecular Properties. In *Applied Multivariate Analysis in SAR and Environmental Studies*; Devilliers, J., Karcher, W., Eds.; ECSC, EEC, EAEC: Brussels and Luxembourg, 1991; pp 247−279.
(40) Downs, G. M.; Willett, P. Clustering in Chemical Structure Databases for Compound Selection. In *Chemometric Methods in Molecular Design*; van der Waterbeemd, H., Ed., VCH: Weinheim, Germany, 1994; pp 111−130.
(41) Downs, G. M.; Willett, P. Similarity Searching in Databases of Chemical Structures. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D., Eds.; VCH Publishers: New York, 1996; Vol. 7; Chapter 1, pp 1−66.
(42) Available on the Internet site with URL http://dtp.nci.nih.gov/.
(43) Ivanciuc, O. Design on Topological Indices. 1. Definition of a Vertex Topological Index in the Case of 4-Trees. *Rev. Roum. Chim.* **1989**, *34*, 1361−1368.
(44) Balaban, T. S.; Filip, P. A.; Ivanciuc, O. Computer Generation of Acyclic Graphs Based on Local Vertex Invariants and Topological Indices. Derived Canonical Labeling and Coding of Trees and Alkanes. *J. Math. Chem.* **1992**, *11*, 79−105.
(45) Ivanciuc, O.; Balaban, T.-S.; Balaban, A. T. Design of Topological Indices. Part 4. Reciprocal Distance Matrix, Related Local Vertex Invariants and Topological Indices. *J. Math. Chem.* **1993**, *12*, 309−318.
(46) Diudea, M. V.; Ivanciuc, O.; Nikolić, S.; Trinajstić, N. Matrices of Reciprocal Distance, Polynomials and Derived Numbers. *MATCH (Commun. Math. Comput. Chem.)* **1997**, *35*, 41−64.
(47) Ivanciuc, O.; Ivanciuc, T.; Balaban A. T. Design of Topological Indices. Part 10. Parameters Based on Electronegativity and Covalent Radius for the Computation of Molecular Graph Descriptors for Heteroatom-Containing Molecules. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 395−401.
(48) Barysz, M.; Jashari, G.; Lall, R. S.; Srivastava V. K.; Trinajstić, N. On the Distance Matrix of Molecules Containing Heteroatoms. In *Chemical Applications of Topology and Graph Theory*; King, R. B. Ed., Elsevier: Amsterdam, 1983; pp 222−227.
(49) Balaban, A. T. Chemical Graphs. Part 48. Topological Index J for Heteroatom-Containing Molecules Taking Into Account Periodicities of Element Properties. *MATCH (Commun. Math. Chem.)* **1986**, *21*, 115−122.
(50) Ivanciuc, O. Design of Topological Indices. Part 12. Parameters for Vertex- and Edge-Weighted Molecular Graphs. *Rev. Roum. Chim.*, in press.
(51) Ivanciuc, O. Design of Topological Indices. Part 19. Computation of Vertex and Molecular Graph Structural Descriptors with Operators. *Rev. Roum. Chim.*, in press.
(52) Ivanciuc, O. Design of Topological Indices. Part 11. Distance-Valency Matrixes and Derived Molecular Graph Descriptors. *Rev. Roum. Chim.* **1999**, *44*, 519−528.
(53) Ivanciuc, O. Design of Topological Indices. Part 14. Distance-Valency Matrixes and Structural Descriptors for Vertex- and Edge-Weighted Molecular Graphs. *Rev. Roum. Chim.*, in press.
(54) Ivanciuc, O.; Balaban, T.-S.; Balaban, A. T. Design of Topological Indices. Part 4. Reciprocal Distance Matrix, Related Local Vertex Invariants and Topological Indices. *J. Math. Chem.* **1993**, *12*, 309−318.
(55) Ivanciuc, O. Design of Topological Indices. Part 16. Matrix Power Operators for Molecular Graphs. *Rev. Roum. Chim.*, in press.
(56) Ivanciuc, O.; Ivanciuc, T. Matrixes and Structural Descriptors Computed from Molecular Graph Distances. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: Amsterdam, The Netherlands, 1999; pp 221−277.
(57) Ivanciuc, O. Design of Topological Indices. Part 18. Modeling the Physical Properties of Alkanes with Molecular Graph Descriptors Derived from the Hosoya Operator. *Rev. Roum. Chim.*, in press.
(58) Sneath, P. The Application of Computers to Taxonomy. *J. Gen. Microbiol.* **1957**, *17*, 201−226.
(59) Sokal, R.; Michener, C. D. A Statistical Method for Evaluating Systematic Relationships. *Univ. Kan. Sci. Bull.* **1958**, *38*, 1409−1438.
(60) Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236−244.
(61) Ward, J. H.; Hook, M. E. Application of an Hierarchical Grouping Procedure to a Problem of Grouping Profiles. *Educ. Psychol. Meas.* **1963**, *23*, 69−82.
(62) Jarvis, R. A.; Patrick, E. A. Clustering Using a Similarity Measure Based on Shared Nearest Neighbors. *IEEE Trans. Comput.* **1973**, *C-22*, 1025−1034.
(63) *BCI Clustering Package, versions 2.5 and 3.0*; Barnard Chemical Information, Ltd.: Sheffield, U.K.; http://www.bci1.demon.co.uk/.
(64) Anderberg, G. M. *Cluster Analysis for Applications*; Academic Press: New York, 1973.
(65) Dubes, R.; Jain, A. K. Clustering Methodologies in Exploratory Data Analysis. *Adv. Comput.* **1980**, *19*, 113−228.
(66) Shenkin, P. S.; Erman, B.; Mastrandrea, L. D. Information-Theoretical Entropy as a Measure of Sequence Variability. *Proteins: Struct. Funct. Genet.* **1991**, *11*, 297−313.
(67) Shenkin, P. S.; McDonald, D. Q. Cluster Analysis of Molecular Conformations. *J. Comput. Chem.* **1994**, *15*, 899−916.
(68) Shannon, C. E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Urbana, IL, 1949.
(69) Duda, R. O.; Hart, P. E. *Pattern Classification and Scene Analysis*; John Wiley and Sons: New York, 1973.
(70) Massart, D. L.; Kaufman, L. *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*; John Wiley and Sons: New York, 1983.
(71) Everitt, B. S. *Cluster Analysis*; Halsted: New York, 1980.
(72) Kaufman, L.; Rousseeuw, P. J. *Finding Groups in Data*; Wiley-Interscience: New York, 1990; Chapter 5, p 238.

CI990149Y