JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

ARTICLE

pubs.acs.org/jcim

Comparison of Combinatorial Clustering Methods on Pharmacological Data Sets Represented by Machine Learning-Selected Real Molecular Descriptors

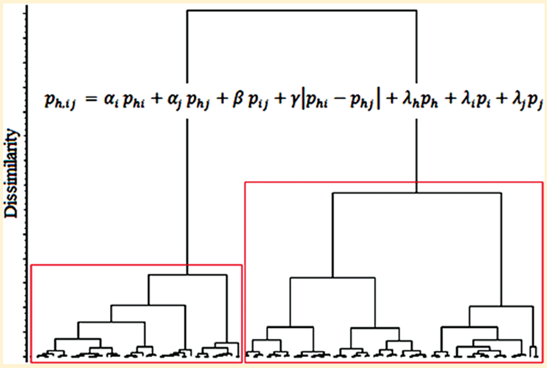
Oscar Miguel Rivera-Borroto,*[†,‡] Yovani Marrero-Ponce,[‡] José Manuel García-de la Vega,[§] and Ricardo del Corazón Grau-Ábalo[†]

[†]Laboratorio de Bioinformática, Centro de Estudios de Informática, Facultad de Matemática, Física y Computación, Universidad Central "Marta Abreu" de Las Villas (UCLV), Santa Clara, 54830 Villa Clara, Cuba

[‡]Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research (CAMD-BIR Unit), Faculty of Chemistry-Pharmacy, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba

[§]Departamento de Química Física Aplicada, Facultad de Ciencias, Universidad Autónoma de Madrid (UAM), 28049 Madrid, Spain

ABSTRACT: Cluster algorithms play an important role in diversity related tasks of modern chemoinformatics, with the widest applications being in pharmaceutical industry drug discovery programs. The performance of these grouping strategies depends on various factors such as molecular representation, mathematical method, algorithmical technique, and statistical distribution of data. For this reason, introduction and comparison of new methods are necessary in order to find the model that best fits the problem at hand. Earlier comparative studies report on Ward's algorithm using fingerprints for molecular description as generally superior in this field. However, problems still remain, i.e., other types of numerical descriptions have been little exploited, current descriptors selection strategy is trial and error-driven, and no previous comparative studies considering a broader domain of the combinatorial methods in grouping chemoinformatic data sets have been conducted. In this work, a comparison between combinatorial methods is performed, with five of them being novel in cheminformatics. The experiments are carried out using eight data sets that are well established and validated in the medical chemistry literature. Each drug data set was represented by real molecular descriptors selected by machine learning techniques, which are consistent with the neighborhood principle. Statistical analysis of the results demonstrates that pharmacological activities of the eight data sets can be modeled with a few of families with 2D and 3D molecular descriptors, avoiding classification problems associated with the presence of nonrelevant features. Three out of five of the proposed cluster algorithms show superior performance over most classical algorithms and are similar (or slightly superior in the most optimistic sense) to Ward's algorithm. The usefulness of these algorithms is also assessed in a comparative experiment to potent QSAR and machine learning classifiers, where they perform similarly in some cases.



INTRODUCTION

Cluster analysis is the organization of a collection of patterns (chemical patterns in this case) into clusters on the bases of similarity. Intuitively, patterns within a valid cluster are more similar to each other than patterns belonging to a different cluster. The clustering process can be divided into the following stages: data collection, initial screening, representation, clustering tendency, clustering strategy, validation, and interpretation.[1] The clustering strategy or grouping step involves a careful choice of the clustering algorithm and initial parameters. *Hierarchical* clustering algorithms produce a nested series of partitions based on criterion for merging or splitting clusters on the bases of similarity. *Partitional* clustering algorithms identify the partition that optimizes a clustering criterion. Additional techniques for the grouping operation include *probabilistic* and *graph-theoretic* clustering methods. Also, the resulting output clustering can be either *hard*, that is, data are partitioned into compact and well separated groups, or *fuzzy*, where each pattern has a variable degree of membership in each of the output clusters.[2] Clustering methods have a long tradition in chemistry, largely due to various groups of scientists working within pharmaceutical companies, where they have to process very large and high-dimensional data sets. Typical applications of their work include high-throughput screening, combinatorial chemistry, compound acquisition, variable reduction, and QSAR.[3]

As a characteristic of grouping methodologies, the resulting clusters depend on the description chosen for chemical patterns and conditioning of the data set, nature of the method, merge criterion, algorithmical technique, and statistical distribution of data.

Received: January 6, 2011
Published: November 18, 2011

dx.doi.org/10.1021/ci2000083 | *J. Chem. Inf. Model.* 2011, 51, 3036–3049

Generally, this last information it is not known a priori and comparative tests among the methods employed are necessary in order to find the model that best fits the problem or family of problems at hand. Since the pioneer works of Adamson and Bush,[4] various comparative studies on the performance of clustering algorithms have been reported in the literature highlighting their relative merits and drawbacks in grouping chemoinformatic data sets.[5] In this direction, results from academic studies and also subsequent applications in pharmaceutical industry programs have confirmed the general superiority of the effectiveness of the *Ward's* method, using the efficient reciprocal nearest-neighbors (RNN) algorithm and 2D fingerprints for vector molecular representation over other hierarchical methods such as *group-average* and *minimum-diameter* and nonhierarchical methods such as *Jarvis—Patrick* and *k-means* in grouping small to medium data sets.[3a,6] Also, novel clustering strategies (mostly graph-based) have been proposed in order to shorten the time of grouping chemical data sets. However, rating their performance has relied more on the efficiency criterion than on their relative effectiveness in respect to the classical methods.[7]

From analysis of the literature, it is apparent that binary strings or *fingerprints* (e.g., MACCs keys, Daylight fingerprints, BCI keys, etc.) have been adopted as the de facto means for molecular representation;[8] the main reason lies in the efficiency with which they can be generated, stored, and compared.[9] Alternatively, some authors have introduced other types of numerical description, extracting more chemical information from molecular entities.[10] Currently, this information can be accessed in a good approximation with software of molecular optimization and descriptor calculation.[11] Although, from the computational point of view, using numerical descriptors is a less efficient strategy than using fingerprints, from the statistical perspective transforming continuous (or discrete) scale descriptors into binary (nominal) scale descriptors leads to a loss of information that affects the power and ability of resolving ties in proximity of the grouping methods.[12] It is also noted that feature selection is an almost ignored stage in experimental studies comparing clustering methodologies on the basis of chemoinformatic data sets despite that most of them have been conducted in supervised scenarios (see for example refs 5a, 5d, 5e, and 10b). However, the importance of the automatic feature selection for cluster analysis has been mentioned in other areas.[13] Molecular descriptor selection has been driven by experience from former studies or by the "trial and error" strategy. As far as we know, few works have used machine learning techniques at this stage;[14] although this not consistent with the *similarity principle* because only linearly dependent descriptors with the response variable satisfy the *neighborhood principle*.[15] Consequently, we also believe that having an a priori rational knowledge based on the behavior of descriptors in supervised contexts may permit better decision making on descriptors selection in analogous but unsupervised contexts. This strategy would be a type of "learning to learn", which is a type of *meta-learning*.[16]

Currently, the Ward's method with the RNN implementation remains the "preferred" algorithm for chemoinformatic grouping tasks. Interestingly, no comparative study has appeared to extend to a more comprehensive range of combinatorial methods. Analogously, because these methods are mathematically and algorithmically similar to Ward's, these should perform similarly in grouping data sets. This research hypothesis has been studied in ecology but not in cheminformatics.[17]

The lack of a published study that treats several combinatorial family members motivated us to present results that cover this apparent gap in the literature. In this direction, the general objective of the present report is to compare combinatorial methods' performance in grouping pharmacological data sets, with the molecular entities being represented by descriptors of a real numerical nature selected by machine learning techniques, according to the neighborhood behavior principle.

MATERIALS AND METHODS

Theoretical Background on Combinatorial Methods. Combinatorial methods are a family of the broader Sequential, Agglomerative, Hierarchical and Nonoverlapping clustering techniques (SAHN methods).[18] They only require a symmetric proximity (dissimilarity, similarity, etc.) matrix **P** to be stored in computer memory during computations; the raw data may be released once this matrix has been calculated (*stored matrix* approach[19]). The original data are not needed because there is a *combinatorial* solution to recompute between cluster measures using the information contained in **P** and in an array of cluster sizes.

The first model describing such behavior was presented by Lance and Williams.[20] They suggested the recurrence formula

$$p_{h,ij} = \alpha_i p_{hi} + \alpha_j p_{hj} + \beta p_{ij} + \gamma |p_{hi} - p_{hj}| \tag{1}$$

to update the values of **P**. That is, if clusters C_i and C_j are merged in a clustering cycle, then $p_{h,ij}$ gives the updated criterion value to be used in the next cycle for merging the cluster $C_i \cup C_j$ with C_h. By setting each combination of parameters $(\alpha_i, \alpha_j, \beta, \gamma)$ to certain constants or functions depending on cluster sizes and substituting the values of p with the characteristic merge criteria, it was possible to reproduce the behavior of seven cluster algorithms known by that time (single linkage, complete linkage, group average, centroid, median, simple average, and incremental sum of squares or Ward's). However, the growing evidence on the combinatorial nature of new methods not fitted to eq. 1 allowed Jambu and Lebeaux[21] to propose the formula

$$p_{h,ij} = \alpha_i p_{hi} + \alpha_j p_{hj} + \beta p_{ij} + \gamma |p_{hi} - p_{hj}| + \lambda_h p_h + \lambda_i p_i + \lambda_j p_j \tag{2}$$

that was able to explain also other seven cluster strategies that appeared later in the literature (minimum increase of error variance, minimum error sum of squares, and minimum error variance of the newly formed clusters, weighted and unweighted average linkage within the new group, and minimum increase of weighted and unweighted average linkage). Interpretation of parameters remained the same, that is, each combination of the type $(\alpha_i, \alpha_j, \beta, \gamma, \lambda_h, \lambda_i, \lambda_j)$ defines a particular algorithm, but now the updated value of $p_{h,ij}$ is determined according to six values of **P**

$$\begin{pmatrix} p_{hh} & p_{hi} & p_{hj} \\ \cdot & p_{ii} & p_{ij} \\ \cdot & \cdot & p_{jj} \end{pmatrix} \tag{3}$$

where, $p_{h,ij} = p[C_h \cup (C_i \cup C_j)]$; $p_{ij} = p(C_i \cup C_j)$; $p_{ii} = p_i = p(C_i)$.

Researchers also generalized eqs. 1 and 2 in two other ways. The first strategy consisted in letting the sets of equation parameters $(\alpha_i, \alpha_j, \beta, \gamma)$, and later $(\alpha_i, \alpha_j, \beta, \gamma, \lambda_h, \lambda_i, \lambda_j)$ correspondingly, to vary continuously, while keeping a fixed functional dependency and certain constrains among them. It allowed disposing of two groups of flexible methods generalizing the "rigid" ones already known. The first group (β-flexible and β,γ-flexible) served to describe transitions between extremely space-dilating and space-contracting algorithms,[20,22] while the

Table 1. Parameters for Combinatorial d-CSAHN (1−6), nh-CSAHN (7−9), and ch-CSAHN (10−12) Clustering Algorithms[17] [a,b]

clustering method	α_i	α_j	β	γ	λ_h	λ_i	λ_j
1. single linkage (SL)	$1/2$	$1/2$	0	$-1/2$	0	0	0
2. complete linkage (CL)	$1/2$	$1/2$	0	$1/2$	0	0	0
3. group average (GA)	$n_i/(n_h + n_i)$	$n_j/(n_h + n_j)$	0	0	0	0	0
4. simple average (SA)	$1/2$	$1/2$	0	0	0	0	0
5. centroid (Cen)	$n_i/(n_h + n_i)$	$n_j/(n_h + n_j)$	$-(n_i n_j)/(n_i + n_j)^2$	0	0	0	0
6. median (Med)	$1/2$	$1/2$	$-(1/4)$	0	0	0	0
7. minimum sum of squares of new cluster (MSSN)	$(n_h + n_i)/n_{.}$	$(n_h + n_j)/n_{.}$	$(n_i + n_j)/n_{.}$	0	$-(n_h/n_{.})$	$-(n_i/n_{.})$	$-(n_j/n_{.})$
8. minimum variance of new cluster (MVN)	$((n_h + n_i)/n_{.})^2$	$((n_h + n_j)/n_{.})^2$	$((n_i + n_j)/n_{.})^2$	0	$-(n_h/n_{.})^2$	$-(n_i/n_{.})^2$	$-(n_j/n_{.})^2$
9. minimum average distance within new cluster (MADN)	$-b_{hi}/b_{.}$	$-b_{hj}/b_{.}$	$-b_{ij}/b_{.}$	0	$-b_h/b_{.}$	$-b_i/b_{.}$	$-b_j/b_{.}$
10. minimum increase of sum of squares or Ward's (MISS)	$(n_h + n_i)/n_{.}$	$(n_h + n_j)/n_{.}$	$-(n_h/n_{.})$	0	0	0	0
11. minimum increase of variance (MIV)	$((n_h + n_i)/n_{.})^2$	$((n_h + n_j)/n_{.})^2$	$-(n_h(n_i + n_j)/n_{.}^2)$	0	0	0	0
12. minimum increase of average distance (MIAD)	$b_{hi}/b_{.}$	$b_{hj}/b_{.}$	$(b_{ij}/b_{.}) - ((b_{ij})/(b_h + b_{ij}))$	0	p_5	p_6	p_7

[a] $b_i = \binom{n_i}{2}$, where n_i is the number of objects in cluster C_i and $n_{.} = n_h + n_i + n_j$. [b] $p_5 = (b_h b_{hi})/((b_h + b_i)) + (b_h b_{hj})/((b_h + b_j)) - (b_h b_{.})/((b_h + b_{ij})) - b_h$, $p_6 = (b_i b_{ij})/((b_i + b_j)) + (b_i b_{hi})/((b_h + b_i)) - (b_i b_{.} b_{ij})/((b_h + b_{ij})(b_i + b_j)) - b_i$, $p_7 = (b_j b_{ij})/((b_i + b_j)) + (b_j b_{hj})/((b_h + b_j)) - (b_j b_{.} b_{ij})/((b_h + b_{ij})(b_i + b_j)) - b_j$.

second one (λ-flexible) also allowed for the monotonicity behavior of its family.[17]

The second strategy focused on the merge criterion of the type $p_{ij} = p(C_i \cup C_j)$, where p depends on the set of pairwise distances between members of cluster $C_i \cup C_j$. In this direction, Batagelj[23] showed that for Ward's-like cluster algorithms, i.e., for which the merge criterion depends on the Gower-Bock dissimilarity,

$$p_{ij} = p(D^G(C_i, C_j)) = p(d(\overline{C}_i, \overline{C}_j)) \qquad (4)$$

where $D^G(C_i,C_j) = d(\overline{C}_i,\overline{C}_j)$ is the Gower-Bock dissimilarity or Euclidean distance d between cluster centers $\overline{C}_i$ and $\overline{C}_j$, the scope of this group of techniques can be extended by considering eq. 4 for d any dissimilarity measure in general rather than just the squared Euclidean distance. With the introduction of the notion of *generalized cluster centers* and *generalized weights*, Batagelj[23] actually mathematically formalized this practice while keeping the notion of *cluster center* that is based on intuition, relying on the Euclidean distance, so eq. 4 can be rewritten as

$$p_{ij} = p(D^G_{\text{gen}}(C_i, C_j)) = p(d(\tilde{C}_i, \tilde{C}_j)) \qquad (5)$$

where $(D^G_{\text{gen}}(C_i,C_j)) = d(\tilde{C}_i,\tilde{C}_j)$ is the generalized Gower-Bock dissimilarity or dissimilarity d between the generalized cluster centers $\tilde{C}_i$, and $\tilde{C}_j$.[23] This means that, for example, in the case of the generalized Ward's method

$$p_{ij} = D^{\text{Ward}}_{\text{gen}}(C_i, C_j) = \frac{w(C_i) \times w(C_j)}{w(C_i \cup C_j)} \times d(\tilde{C}_i, \tilde{C}_j) \qquad (6)$$

the factor $d(\tilde{C}_i,\tilde{C}_j)$, where $w(C_i)$ is the generalized weight of the cluster C_j and so on, does not need to be necessarily the traditional squared Euclidean distance (Hamming distance for binary vectors) but can also be any other dissimilarity measure such as the Soergel distance (complement of Tanimoto coefficient).[24]

In order to organize the experience in this field, Podani suggested a novel classification of combinatorial algorithms based on the original ideas of Lance and Williams, depending on the nature of the dissimilarity measure that each method use as the merge criterion:

1. d-CSAHN: The merge criterion is given as an intercluster dissimilarity measure, i.e., the entries of **P** are given as $p_{ij} = d(C_i, C_j)$ or $p_{ij} = d^2(C_i, C_j)$.

2.1. nh-CSAHN: The merge criterion is given as an intracluster homogeneity measure, i.e., the entries of **P** are given as $p_{ij} = h(C_i, C_j)$.

2.2. ch-CSAHN: The merge criterion is given as a change in the homogeneity measure, i.e., the entries of **P** are given as $p_{ij} = h(C_i, C_j) - h(C_i) - h(C_j)$.

As agglomeration proceeds, criteria of type 1 and 2.2 are minimized, while criteria of type 2.1 are maximized.[17]

Efficiency of Algorithms. There are two basic algorithmical techniques to perform the matrix **P** update, *closest pair* and *reciprocal nearest neighbors*.[25] Although the second technique improves the first one, as pertains to the required resources [$O(N^3)$ and $O(N^2)$ vs $O(N^2)$ and $O(N)$ for expected time and space complexities, respectively], it also affects the ultrametric properties of some methods, namely, what has been noted as undesired *reversals* in the respective dendrograms.[17]

In this work, we compared five novel cluster algorithms in cheminformatics, minimum sum of squares of new cluster (MSSN), minimum variance of new cluster (MVN), minimum average distance within new cluster (MADN), minimum increase of variance (MIV), and minimum increase of average distance (MIAD), to the Ward's method (MISS) and six other classical grouping methods. Table 1 shows a summary of these combinatorial methods, with their corresponding parameters referred to eq. 2.

Datasets Domain for Methods Validation. The performance of similarity measures, molecular descriptors, and even validation approaches is strictly dependent on the test molecules/databases, chemical space configuration, and problematic treated. No *standard* testing data set has been adopted for the community, probably due to the impossibility in finding a unique group of molecules that regroups all the screening needs of today's cheminformatics.[26] For this reason, it has been suggested that for validation purposes investigators present at least 10 examples with diverse activities and that there be more than one standard of comparison.[27]

In this study, eight different medicinal chemistry data sets were taken from the original work of Sutherland et al.[28] These chemical

Table 2. Description of Medicinal Chemistry Data Sets Used in This Study

data set[a]	pharmacological target	number of compounds	pharmacokinetic variable[b]	range of values
ACE	angiotensin converting enzyme inhibitors	114	pIC50	2.1−9.9
AchE	acetyl-cholinesterase inhibitors	111	pIC50	4.3−9.5
BZR	ligands for the benzodiazepine receptor	163	pIC50	5.5−8.9
COX-2	cyclooxygenase-2 inhibitors	322	pIC50	4.0−9.0
DHFR	dihydrofolate reductase inhibitors	397	pIC50	3.3−9.8
GPB	glycogen phosphorylase b inhibitors	66	pK_i	1.3−6.8
THER	thermolysin inhibitors	76	pK_i	0.5−10.2
THR	thrombin inhibitors	88	pK_i	4.4−8.5

[a] Data sets are presented in the order of the original source.[28] These are freely available at http://www.cheminformatics.org/data sets/index.shtml. [b] pIC50 = −log IC50, where IC50 is half of maximum inhibitory concentration, and it is used as a measure of the drug potency. pK_i = −log K_i, where K_i is the inhibition constant of the drug, also it is used as a measure of the drug potency.

repositories have also been used by other researchers in QSAR studies.[29] A brief description of these data sets is shown in Table 2.

Chemical Space and Molecular Representation. Closely allied with the notion of molecular similarity is that of a *chemical space*. It provides a means for conceptualizing and visualizing molecular similarity. A chemical space consists of a set of molecules and a set of associated relations (e.g., similarities, dissimilarities, or distances) among the molecules, which give the space a "structure".[30]

Chemical space can be described by using a *coordinate-based* coding or a *coordinate-free* coding of molecular structures. In the single molecule coding (coordinate-based space), each molecule is described by a substructure or fragment vector (e.g., count fingerprints) and, therefore, has an absolute position in a multidimensional space whose dimension is specified by the number of uncorrelated features; whereas in pairwise molecule coding (coordinate-free space), only the distances between two molecules are computed, using an explicit or implicit similarity measure (e.g., molecular shape-based methods). The absolute position of molecules in this space can only be calculated by measuring all pairwise distances, and the dimensionality of the space may be unknown.[31]

A very large number of descriptors that can be used in similarity calculations have been developed. They are typically designed to provide a molecular description that is transferable, in an information-preserving representation, to an abstract descriptor space.[32] However, as the dimensionality of the data increases, many types of data analysis and classification problems become significantly more difficult. Sometimes the data also become increasingly sparse in the space they occupy. This can lead to big problems for both supervised and unsupervised learning. In the literature, this phenomenon is referred to as the *curse of dimensionality*.[33] For clustering purposes, the most relevant aspect of the curse of dimensionality concerns the effect of increasing dimensionality on distance or similarity. For certain data distributions, the relative difference in the distances of the closest and farthest data points $(d_{\max} - d_{\min})/(d_{\min})$ of an independently selected point approximates zero as the dimensionality increases. Also, the behavior of the absolute difference between the distance to the closest and farthest neighbors of an independently selected point depends on the distance measure itself. In particular, for the Minkowski's L_1 metric, $d_{\max} - d_{\min}$ (Manhattan distance) increases with dimensionality, for the L_2 metric (Euclidean distance), $d_{\max} - d_{\min}$ remains relatively constant, and for the L_r metric, $r \geq 3$, $d_{\max} - d_{\min}$ approximates zero as dimensionality increases. These theoretical results have also been confirmed by experiments on simulated and real data sets.[34] Sometimes, a large number of features or descriptors may contain irrelevant or weakly relevant features that negatively affect the accuracy of prediction algorithms.[35] The extreme case of this phenomenon is depicted in the Watanabe's *ugly duckling theorem*. Basically, if one considers the universe of an objects features and has no preconceived bias about which features are better, no matter which pair of objects one compares, all will be equally similar (dissimilar).[36] One solution to this problem is to select a particular set of descriptors for which good performance in a certain problem was shown. A further strategy is to first calculate a large number of descriptor values and later remove those descriptors from the set, which shows a correlation coefficient above a certain value. A different approach is to let the computer choose the optimal combination of descriptors for a given problem.[37]

In our study, molecules were represented as real vectors of the type $m = (I_1, I_2, I_3, ..., I_n)$, where the values for I_k are the molecular indices. A collection of such vectors constituted the data set matrix **M**, which after an appropriate conditioning was suitable for cluster analysis. The description for obtaining the final numerical data sets is as follows.

Data sets were handled with the JChem for Excel utility.[38] Each was reoptimized with the 3D structure generator software CORINA.[39] The most relevant software parameters fixed in this process were wh, write added hydrogen atoms; rs, remove small fragments; and neu, neutralize formal charges. Output files were loaded in the software for molecular descriptors calculation DRAGON.[40] In this stage, all molecular descriptors families available were computed (a total of 3224 descriptors), and then binary features were removed. Resulting files were loaded into the software of data mining Weka[41] and then were subjected to a treatment including prefiltration, rescaling, and feature selection processes. At this stage, nominal attributes were removed with the *RemoveType* filter; attributes that do not vary at all or that vary too much were removed with the *RemoveUseless* filter, while keeping default parameters. The resulting numerical attributes as well as the class attribute were standardized to have zero mean and unit variance with the *Standardize* filter. Feature selection was performed by using the *AtributeSelection* filter. Here, *CfsSubsetEval* was set with the rest of default parameters, and it selects subsets of features that are highly correlated with the class, while having low intercorrelation among them.[42] Our criterion is that a supervised correlation-based subset evaluator is important for cluster analysis of chemoinformatic data sets because it warrants that similar chemical structures group in the same cluster, whereas dissimilar molecules group in different clusters. Only linearly dependent descriptors with the class satisfy the *neighborhood principle*.[15b] Additionally, we choose the *CfsSubsetEval* evaluator because of its simplicity and because it has been used by other researchers on the same data sets, obtaining relatively good results in the accuracy scores of the compared classifiers.[29c,d]

As a final step, resulting files were loaded in the software for data analysis on ecology and systematics SYN-TAX2000.[43] In this stage, some available options were fixed for the entire process: *Euclidean distance* as the proximity measure, *suboptimal fusion* technique for ties resolution,[44] and the *closest pair* as the fusion strategy.[25] After running each cluster algorithm, the corresponding tree was pruned at a height of two branches (clusters) with the software facility *partition from dendrogram*. This partition of data was first proposed in a QSAR study on a comparison machine of learning classifiers based on the balanced distribution of the numerical class.[29a] Also, it agrees with the Podani's criterion for the optimum number of clusters for the best performing algorithms.[45] The external quality of clusters was assessed through four measures: F-measure, global accuracy, Matthews' correlation coefficient, and Tanimoto's coefficient. The first index has been used traditionally as a subjective measure of clustering quality,[46] next two measures have been applied to assess the accuracy of prediction algorithms,[47] the last one, however, has been successfully used in the virtual screening of chemical data sets with molecular fingerprints.[48]

Statistical Experimental Design. In the literature, grouping methodologies are traditionally handled as unsupervised learning algorithms. In this work, however, with the introduction of a supervised filter for feature selection, they can be assessed in supervised learning settings (akin a QSAR classifier). In this direction, two methodological principles aiding for a rational judgment on comparison are the *no free lunch theorem* (NFL) and *minimum description length principle* (MDL). In order to define them, let us introduce the terms *generalization error* (c); *target input—output relationships* (f), from which *m-elements training sets* (d) are produced; *hypotheses* (h), the outputs of one's learning algorithm made in response to d or alternatively the algorithm's guess for f; and *prior probability not conditioned by the data* [$P(f)$]. Basically, the machine learning NFL theorem states that all learning algorithms are the same in that by several definitions of "average" they have the same average off-training/on-training set misclassification risk (expected value $E[C|.]$) and, therefore, no learning algorithm can have a lower risk than another one for all f, for all $P(f)$, for all f and d, or for all $P(f)$ and d.[49] The MDL principle is a method for inductive inference that provides a generic solution to the model selection problem. It is based on the following insight: any *regularity* in the data can be used to *compress* the data, that is, to describe it using fewer symbols than the number of symbols needed to describe the data literally. MDL combines these two insights by *viewing learning as data compression*: it tells us that for a given set of hypotheses h and data set d, we should try to find the hypothesis or combination of hypotheses in h that compresses d most. The best point hypothesis h to explain the data d is the one which minimizes the sum $L(h) + L(d|h)$, which is a direct measure of complexity, where $L(h)$ is the length, in bits, of the description of the hypothesis, and $L(d|h)$ is the length, in bits, of the description of the data when encoded with the help of the hypothesis.

Although the NFL theorem predicts that comparison becomes useless when results are averaged over the population of chemoinformatic data sets, i.e., data sets are generated by a variety of underlying input—output relationships f (physical, physicochemical, biochemical/pharmacological, etc.) with a given $P(f)$, resulting in different numbers of d, the MDL principle (embodying the Ockham's razor) suggests that it is possible to find the predictive model with the best balance between fitness and simplicity for the studied pharmacological data sets. The reason is that NFL operates on the complete domain of all possible hypotheses, while MDL tries to generalize its hypothesis built from the observed data by limiting its complexity accordingly.

In order to test the relative performance of cluster algorithms, the effectiveness measure on the eight data sets for each of the twelve grouping methods was recorded. The resulting matrix **E** (8×12) can be considered as consisting in twelve samples of effectiveness scores of volume eight, which are matched (that is, cross-classified or stratified, with each stratum contributing one observation to each sample) according to each considered pharmacological activity.[50] After a Friedman's-like or RT-2 transformation,[51] the new matrix $\mathbf{E_r}$ is then suitable for a Friedman's two-way criteria ANOVA.[52]

RESULTS AND ANALYSIS

Behavior of Molecular Descriptors. Linearly relevant descriptors corresponding to each pharmacological activity (medicinal chemistry data set) selected by the Weka *CfsSubsetEval* evaluator are shown in Table A1 in the Appendix. The percentage of dimensionality reduction for binary descriptors removal (Weka selection) stage corresponding to each data set was ACE 55.18% (98.27%), AchE 56.27% (97.94%), BZR 54.28% (98.98%), COX-2 52.70% (98.62%), DHRF 53.07 (98.94%), GBP 55.89% (99.30%), THERM 56.51% (98.93%), and THR 56.36% (98.93%), with the geometric mean of these values being 55.01% (98.74%). These results indicate that approximately 55% of molecular descriptors of DRAGON software are binary and, thus, discarded in this study. Subsequent steps of cleansing and selection in Weka yielded 98.74% of nonrelevant features removal. This significant value suggests a high degree of specificity in molecular features—pharmacological activity relationships.

In order to study the statistical behavior of selected features, the empirical frequency distribution (*efd*) of the categorical variable "family" corresponding to each data set was compared to the *efd* of this variable for the fusion data set (i.e., the eight data sets are considered as one data set) and also to the a priori *efd* of DRAGON. To this end, a χ^2 goodness of fit test was performed (Table 3).

From the probability scores presented in the previous table, one can infer a significant similarity between each specific data set *efd* and the fusion data set *efd* with respect to the variable "family", except for the DHFR repository. In other words, a few descriptor families seem to capture the chemical information necessary to describe ligand—receptor interaction phenomena latent in the data sets domains. This tendency is visualized in Chart 1a and b for "family" and "dimensionality" variables, respectively, considering the fusion data set as reference.

Moreover, it is interesting to note that chemical information codified by 2D and 3D descriptors is closely related to pharmacological activities; they contribute 84% out of selected features. A more refined comparison of proportions confirmed that contributions of these two groups are significatively greater ($p \sim 0$ in both cases) than the null hypothesis probability ($p = 1/6$). A similar analysis by considering the descriptor families showed (see Table 4) that contributions of 2D autocorrelations (2D), RDF descriptors (3D), 3D-MoRSE descriptors (3D), and WHIM descriptors (3D) families are significantly greater than the null hypothesis probability ($p = 1/16$), which is consistent with the

Table 3. Significance of the χ^2 Goodness of Fit Test between Empirical Frequency Distributions of Molecular Descriptors

reference[a]	fusion	ACE	AchE	BZR	COX-2	DHFR	GBP	THERM	THR
fusion	1	0.4300	0.5770	0.7245	0.8601	0.0081**	0.7834	0.8080	0.2837
DRAGON	∼ 0***	0.0011**	∼ 0***	0.0022**	0.0027**	0.0031**	0.0152*	0.1840	0.0045**

[a] Reference distributions for comparison. * Significant statistical tests ($p < 0.05$). ** Highly significant statistical tests ($p < 0.01$). *** Extremely significant statistical tests ($p < 0.001$).

Chart 1. (a) Relative contribution of descriptors to the fusion data set according to the *family* criterion. (b) Relative contribution of descriptors to the fusion data set according to the *dimensionality* criterion.

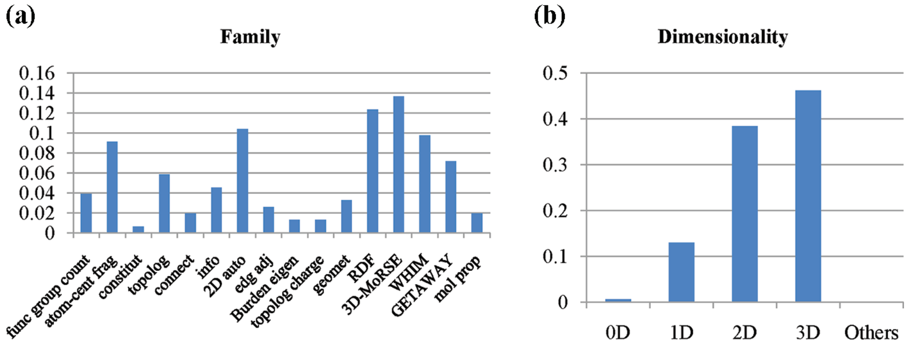


Table 4. Binomial Tests for Homogeneity in Descriptor Families

family	sig[a]	family	sig[a]	family	sig[a]
functional group counts	0.8499	2D autocorrelations	**0.0112****	3D-MoRSE descriptors	**∼ 0****
atom-centered fragments	0.0528	edge adjacency indices	0.9554	WHIM descriptors	**0.0256***
constitutional descriptors	0.9965	Burden eigenvalues	0.9909	GETAWAY descriptors	0.2666
topological descriptors	0.5161	topological charge indices	0.9909	molecular properties	0.9789
connectivity indices	0.9789	geometrical descriptors	0.9143		
information indices	0.7593	RDF descriptors	**0.0005*****		

[a] Statistical significance, work alternative hypothesis $p > 0.0625$ (1/16). The normal approximation was used with the classical correction $X + 0.5$ for the random variable. * Significant statistical tests ($p < 0.05$). ** Highly significant statistical tests ($p < 0.01$). *** Extremely significant statistical tests ($p < 0.001$).

previous result. Briefly, 2D autocorrelations descriptors are based on the autocorrelation function and describe how a considered property is distributed along a topological molecular structure. RDF (*radial distribution function*) descriptors are based on a radial distribution function that can be interpreted as the probability distribution of finding an atom in a spherical volume of radius *R*. 3D-MoRSE (*3D-molecule representation of structures based on electron diffraction*) descriptors are based on the idea of obtaining information from the 3D atomic coordinates by the transformation used in electron diffraction studies for preparing theoretical scattering curves. Lastly, WHIM descriptors (*weighted holistic invariant molecular descriptors*) are geometrical descriptors based on statistical indices calculated on the projections of the atoms along principal axes. They are built in such a way as to code relevant molecular 3D information regarding molecular size, shape, symmetry, and atom distribution with respect to invariant reference frames.[11,40]

On theoretical chemistry grounds, both results suggest that spatial configuration of atoms as well as molecular topology play a dominant role in ligands−receptor interactions. This agrees with similar regularities reported by other authors.[27]

Performance of Clustering Methods. A preliminary analysis based on the visual inspection of dendrograms indicates some regularity in the behavior of clustering methods (see for example Figure 1). Methods SL and MIAD show a strong degree of chaining in all of their agglomeration trees; GA and SA do not retrieve properly the data structure in the form of two underlying clusters, except SA for the ACE data set; Cen and Med show dramatic reversals or monotonicity failures in their dendrograms; MIV shows some degree of chaining and form reversals in most data sets, except for GBP where a good resolution in two clusters was observed. On the other hand, methods CL, MSSN, MVN, MADN, and MISS are able to resolve each chemical repository in two groups. As a general tendency here, at relative short distances, all chemical patterns have merged and distributed into these groups, which later merge at a relatively larger distance. The presence of reversals in Cen and Med can be theoretically explained for these methods that fail to fulfill monotonicity conditions for algorithms of the type d-CSAHN.[53] Similarly, models MIV and MIAD fail to fulfill monotonicity conditions for algorithms of the type h-CSAHN.[17,54] Lastly, the data structured in two clusters provided by the well-behaved algorithms serves as a posteriori support for the dichotomization of pharmacokinetic variables pIC50 and pK_I done in earlier studies.[29a]

In order to quantitatively support the previous analysis, the external quality of groupings was assessed through the

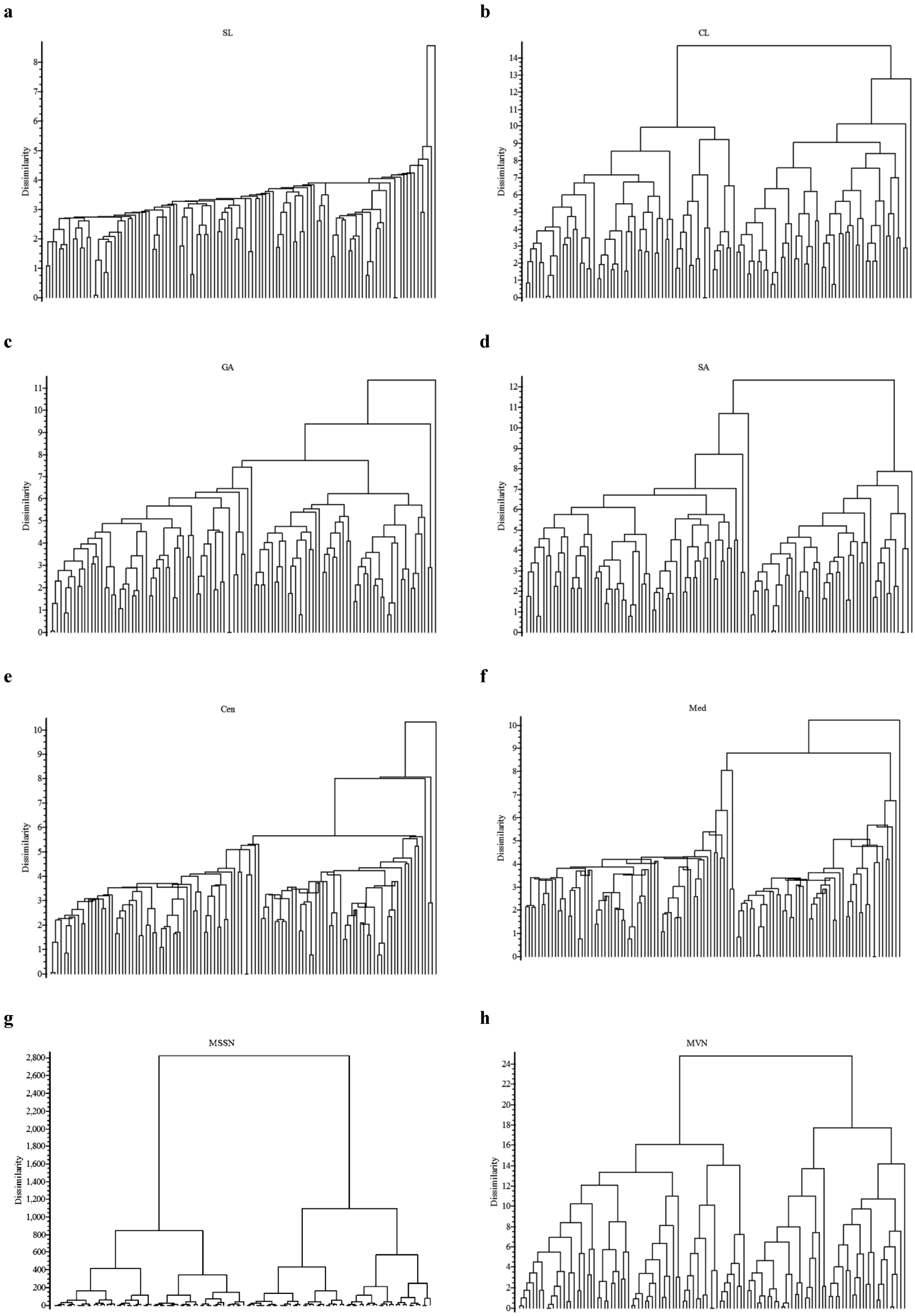


Figure 1. Continued

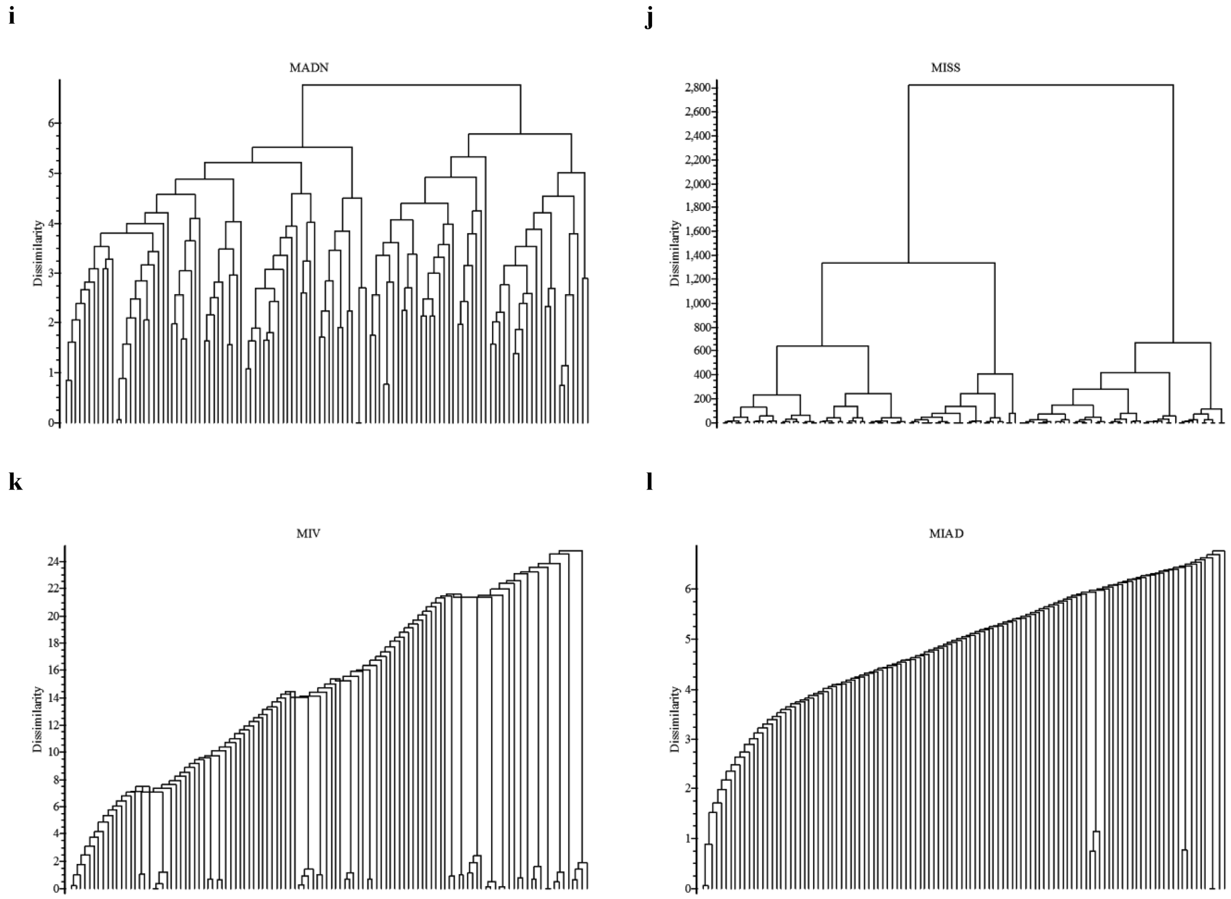


Figure 1. (a−l) Dendrograms for the visual assessment of CSAHN cluster algorithms performance. ACE data set is taken as example. Large groups are represented left.

F-measure (F), global accuracy (Q_t), Matthews' correlation coefficient (C), and Tanimoto's coefficient (T). For our peculiar problem, "assessing the external quality of clustering" can be reformulated as "assessing the accuracy of a binary classifier". In this direction, we provide some definitions based on the information related to this subject.[47,55] Let **O** be a binary vector of dimension n (data set size) representing the external or "natural" classification observed for the data set molecules into active "1" and inactive "0", and let **P** be a binary vector with the same dimension as **O** but representing the cluster classification predicted for those data set molecules into active "1" and inactive "0". Later, formulas corresponding to above coefficients can be uniformly presented as

$$F = \frac{2\mathbf{OP}}{\mathbf{O}^2 + \mathbf{P}^2} \tag{7}$$

$$Q_t = 1 - \frac{(\mathbf{O} - \mathbf{P})^2}{\max\{(\mathbf{O} - \mathbf{P})^2\}} \tag{8}$$

$$C = \frac{(\mathbf{O} - o\mathbf{1})(\mathbf{P} - p\mathbf{1})}{\sqrt{(\mathbf{O} - o\mathbf{1})^2(\mathbf{P} - p\mathbf{1})^2}} \tag{9}$$

$$T = \frac{\mathbf{OP}}{\mathbf{O}^2 + \mathbf{P}^2 - \mathbf{OP}} \tag{10}$$

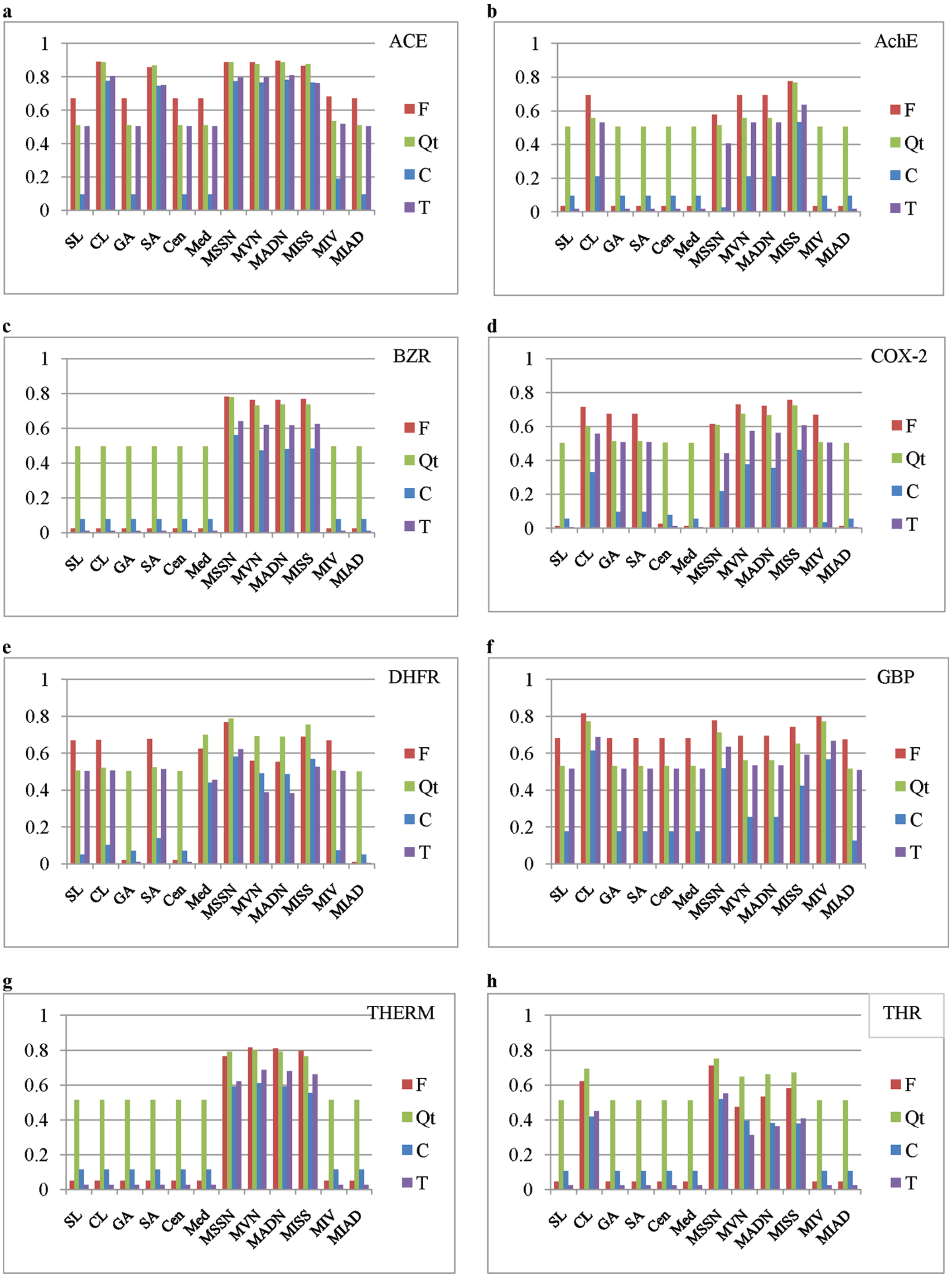
Formulas have been stated in terms of scalar products of vectors; o and p represent the proportion of 1s in **O** and **P**, respectively; **1** represents a vector of 1s. These coefficients were calculated and plotted for each data set as shown in Chart 2.

As can be viewed in this chart, Q_t and C describe in a stable and coherent manner the tendency observed in the former visual analysis. Through all data sets, they assign high values for the algorithms that present good behavior (CL, MSSN, MVN, MADN, and MIV), while rating with low scores to the worse algorithms (SL, GA, SA, Cen, Med, MIV, and MIAD). A value of Q_t above 0.5 indicates a nonrandom assignation of cases to clusters, while the significance of C can be directly evaluated by its relationship to chi-squared statistics as $\chi_1{}^2 = C^2 n$, where n is the data set volume. On the contrary, F and T behave in an unstable manner because they correctly assign low scores for SL, GA, SA, Cen, Med, MIV, and MIAD algorithms in data sets AchE, BZR, and THERM, THR, but they also tend to overrate these algorithms in data sets ACE, COX-2, DHFR, and GBP.

On the basis of previous results, we decided to perform a multiple comparison among the five first methods by using Q_t as a random variable and the two-way ANOVA of Friedman (see Data I in Table 5).

From the Friedman's test probability for Data I, one can infer there are not significant differences in the corresponding median values of Q_t scores for CSAHN methods on studied data sets

Chart 2. (a−h) Stability of external validity measures at assessing the quality of data sets clustering.



($p > 0.05$). Thus, three out of five methods proposed in this work (MSSN, MVN, MADN) perform superiorly to the five classical methods SL, GA, SA, Cen, and Med and similarly (or superiorly in the most optimistic sense for MSSN) to CL and Ward's (MISS). Since "winner" methods are statistically indistinguishable as for accuracy in the current data set domain, CL is preferable because it is the simplest among them, which is consistent with the MDL principle, i.e., its merge formulas have the lowest number of arithmetical operations to perform. However, from a medicinal perspective and counting on considerable computational resources, one can be tempted to choose MSSN in situations where MSSN is marginally superior to the other cluster algorithms (Table 5); for example, for the ligands of the benzodiazepine site (BZR) MSSN, it is plausible to provide at least an additional drug per 100 candidates to reduce anxiety in patients with psychological treatment.

In a further step, the best five algorithms' Q_t scores were averaged for a performance comparison with the other three reported methods, from QSAR and machine learning communities, on the

Table 5. Data and Results for Multiple Comparisons among the Best CSAHN Strategies and for Their Comparison to QSAR and Machine Learning Algorithms

	Data I[a]					Data II[b]			
data sets	CL	MSSN	MVN	MADN	MISS	CSAHN[c]	Bruce[d]	Johansson[e]	Sönströd[f]
ACE	88.60	88.60	87.72	88.60	87.72	88.25	87.84	90.12	84.53
AchE	55.86	51.35	55.86	55.86	76.58	59.10	73.33	70.22	64.80
BZR	49.69	77.91	73.01	73.62	73.62	69.57	76.44	75.56	72.17
COX-2	59.94	60.87	67.39	66.46	72.36	65.40	75.30	74.92	70.30
DHFR	52.14	78.59	69.27	69.02	75.57	68.92	82.17	80.67	76.50
GBP	77.27	71.21	56.06	56.06	65.15	65.15	74.47	76.07	65.40
THERM	51.32	78.95	80.26	78.95	76.32	73.16	70.37	69.57	64.40
THR	69.32	75.00	64.77	65.91	67.05	68.41	68.77	72.90	61.67
Fried[g]			0.481					**0.002****	
Av. Rank[h]	2.50	3.69	2.63	2.81	3.38	1.75	3.38	1.63	3.25
Wilcox[i]		CSAHN-Bruce		**0.039***				Bruce-Sönströd	**0.004****
		CSAHN -Sönströd		0.473				Bruce-Johansson	0.473
		CSAHN -Johansson		**0.012***				Sönströd-Johansson	**0.004****

[a] Q_t values for multiple comparison among the best CSAHN methods. [b] Q_t values for multiple comparison among the CSAHN strategies, and QSAR and machine learning methods reported by other authors. [c] The average value of CSAHN methods for each data set (through each file) was taken. [d] Average of Q_t values reported by this author for tested methods on the same data sets: tree, bagged tree, boosted tree, random forest, SVM, tunned forest, and tunned SVM.[29a] [e] Average of Q_t values reported by this author for tested methods on the same data sets: MLP, RBF, SVM, Bag-M_W, Bag-RBF, Bag-M_B, Avg-M_A, GAS, and NB neural networks.[29c] [f] Average of Q_t values reported by this author for tested methods on the same data sets: RIPPER o JRip (decision lists), C4.5 o J48 (decision tres), and Chipper (decision lists).[29d] [h] Exact signification for the multiple comparison Friedman's χ^2 test. [g] Average rank assigned to each algorithm by the Friedman's test; the higher this value the better the algorithm. [i] Exact signification for the post hoc pairwise comparison one-tailed Wilcoxon's test. * Significant statistical test ($p < 0.05$). ** Highly significant statistical test ($p < 0.01$).

same data sets (see Data II in Table 5). The analogous Friedman result allows for inferring that there are significant differences in the corresponding median values of Q_t scores among these methods ($p < 0.01$). A further step in the analysis would be to find the outlying groups of methods causing such differences; in this sense, the Wilcoxon's signed-rank test can be used to achieve this goal. This contrast test is used to compare two *matched samples* to assess whether their population medians differ.[12] For our case, it is used to compare two lists of quality scores from two different methods on the unique sample of eight data sets and check if there exists a significant difference between them. Results from applying the post hoc Wilcoxon's test (see lower part of Table 5) reveals the presence of two homogeneous groups, CSAHN−Sönströd and Bruce−Johansson, having a nonsignificant intragroup probability of $p > 0.05$, while having a significant intergroup pairwise probability of $p < 0.01$. This means that in median scores CSAHN algorithms perform comparably to simple *Decision Trees* and *Decision List* classifiers, while they are outperformed by more refined multiclassification techniques using *Decision Trees*, *Support Vector Machine*, and *Neural Networks* methods. These two last are considered as potent and promising machine learning techniques in supervised chemoinformatic problems.[56] Because reference models were constructed by their authors using descriptors proposed by Sutherland et al. (probably represented in ours), comparison judgments should be more appropriately done by considering the representation of molecules, or in other words, in the context of the representation method. For this reason, results discussed earlier represent valid evidence not only to support the quality of the best CSAHN algorithms proposed in this work but also to confirm the suitability of the proposed representation for molecules and data sets.

■ CONCLUSIONS

Though our study is limited to a particular group of relatively small data sets, our results suggest that some descriptor families are capable of describing not only specific ligands−pharmacological target interactions but also more general molecular features−biological context relationships. Therefore, in absence of previous knowledge, they may serve as a guide for representing unsupervised chemoinformatic data sets intended for similar medicinal chemistry applications. Moreover, results on the relative performance of clustering algorithms are encouraging because they provide three mathematically and algorithmically well-behaved grouping methods in chemoinformatic tasks, that is, MSSN, MVN, and MADN with a similar yield as the "choice" or Ward's (MISS) algorithm, with MSSN being arithmetically simpler and thus more parsimonious among these three novel algorithms. Comparison of the best five clustering methods with advanced QSAR and machine learning techniques indicates, on median scores, that they perform similarly to supervised classifiers using *Decision Trees* and *Decision Lists*, but they are outperformed by ensembles using these functions and the potent SVM and NN. This suggests that with the use of considerably informative molecular descriptors and a proper descriptor selection technique consistent with the neighborhood principle the power of these combinatorial clustering methods can be increased in a way even comparable to the more refined supervised techniques in their own working arena. In this direction, our future efforts are focused on including novel cluster methodologies for comparison, using a broader domain of large cheminformatic and statistically simulated data sets represented by adequately informative molecular descriptors being consistent with the underlying theory of related classification problems.

APPENDIX

Table A1. List of Molecular Descriptors Selected by the Weka Feature Selection Filter CfsSubsetEval

data	descriptor	family[a]	dim[b]	data	descriptor	family[a]	dim[b]
ACE	MAXDP	topological descriptors	2D	COX-2	Lop	topological descriptors	2D
	PW4	topological descriptors	2D		D/Dr09	topological descriptors	2D
	Lop	topological descriptors	2D		X1A	connectivity indices	2D
	BIC5	information indices	2D		G(N..O)	geometrical descriptors	3D
	ATS4m	2D autocorrelations	2D		RDF060m	RDF descriptors	3D
	MATS8m	2D autocorrelations	2D		RDF060v	RDF descriptors	3D
	MATS3p	2D autocorrelations	2D		Mor12u	3D-MoRSE descriptors	3D
	EEig03d	edge adjacency indices	2D		Mor30m	3D-MoRSE descriptors	3D
	EEig11d	edge adjacency indices	2D		Mor08v	3D-MoRSE descriptors	3D
	EEig12d	edge adjacency indices	2D		Mor30v	3D-MoRSE descriptors	3D
	DISPp	geometrical descriptors	3D		Mor12e	3D-MoRSE descriptors	3D
	RDF035u	RDF descriptors	3D		E3u	WHIM descriptors	3D
	RDF035m	RDF descriptors	3D		P1v	WHIM descriptors	3D
	RDF035e	RDF descriptors	3D		E1e	WHIM descriptors	3D
	RDF035p	RDF descriptors	3D		R6u+	GETAWAY descriptors	3D
	Mor23m	3D-MoRSE descriptors	3D		*R3m+*	GETAWAY descriptors	3D
	Mor26v	3D-MoRSE descriptors	3D		H-049	atom-centered fragments	1D
	Mor26p	3D-MoRSE descriptors	3D		O-058	atom-centered fragments	1D
	E3u	WHIM descriptors	3D		F03[N−O]	2D frequency fingerprints	2D
	E1p	WHIM descriptors	3D		F05[N−N]	2D frequency fingerprints	2D
	C-006	atom-centered fragments	1D		F07[N−F]	2D frequency fingerprints	2D
	C-026	atom-centered fragments	1D	DHFR	nR05	constitutional descriptors	0D
	ALOGP2	molecular properties	Others		D/Dr10	topological descriptors	2D
	F03[O−O]	2D frequency fingerprints	2D		GATS7m	2D autocorrelations	2D
	F06[O−O]	2D frequency fingerprints	2D		GATS6p	2D autocorrelations	2D
AchE	D/Dr07	topological descriptors	2D		BELm2	Burden eigenvalues	2D
	IC4	information indices	2D		BELe1	Burden eigenvalues	2D
	SIC5	information indices	2D		RCI	geometrical descriptors	3D
	BIC5	information indices	2D		Mor10u	3D-MoRSE descriptors	3D
	MATS4m	2D autocorrelations	2D		Mor03m	3D-MoRSE descriptors	3D
	MATS4p	2D autocorrelations	2D		Mor04m	3D-MoRSE descriptors	3D
	GATS4m	2D autocorrelations	2D		Mor09e	3D-MoRSE descriptors	3D
	GATS6e	2D autocorrelations	2D		R5u	GETAWAY descriptors	3D
	GATS5p	2D autocorrelations	2D		C-033	atom-centered fragments	1D
	JGI10	topological charge indices	2D		O-057	atom-centered fragments	1D
	RDF045u	RDF descriptors	3D		F04[C−N]	2D frequency fingerprints	2D
	RDF090u	RDF descriptors	3D		F04[N−O]	2D frequency fingerprints	2D
	RDF155u	RDF descriptors	3D	GBP	X5A	connectivity indices	2D
	RDF090m	RDF descriptors	3D		BIC1	information indices	2D
	RDF090e	RDF descriptors	3D		MATS8v	2D autocorrelations	2D
	RDF155e	RDF descriptors	3D		MATS7e	2D autocorrelations	2D
	Mor22m	3D-MoRSE descriptors	3D		Mor13m	3D-MoRSE descriptors	3D
	Mor11e	3D-MoRSE descriptors	3D		R5m+	GETAWAY descriptors	3D
	Mor32e	3D-MoRSE descriptors	3D		C-006	atom-centered fragments	1D
	E3u	WHIM descriptors	3D		H-046	atom-centered fragments	1D
	G2m	WHIM descriptors	3D		F02[N−O]	2D frequency fingerprints	2D
	G3v	WHIM descriptors	3D		F07[O−O]	2D frequency fingerprints	2D

Table A1. Continued

data	descriptor	family[a]	dim[b]	data	descriptor	family[a]	dim[b]
	G1e	WHIM descriptors	3D	THERM	X5v	connectivity indices	2D
	E3p	WHIM descriptors	3D		IC1	information indices	2D
	R6e+	GETAWAY descriptors	3D		GATS5m	2D autocorrelations	2D
	nR=Cs	functional group counts	1D		GATS7p	2D autocorrelations	2D
	nArCONR2	functional group counts	1D		RDF065m	RDF descriptors	3D
	H-053	atom-centered fragments	1D		Mor17m	3D-MoRSE descriptors	3D
	O-058	atom-centered fragments	1D		Mor31m	3D-MoRSE descriptors	3D
BZR	TI2	topological descriptors	2D		Mor16e	3D-MoRSE descriptors	3D
	Vindex	information indices	2D		Du	WHIM descriptors	3D
	ATS7e	2D autocorrelations	2D		R5p	GETAWAY descriptors	3D
	J3D	geometrical descriptors	3D		nCt	functional group counts	1D
	HOMA	geometrical descriptors	3D		nROH	functional group counts	1D
	RDF020u	RDF descriptors	3D		F01[O−S]	2D frequency fingerprints	2D
	RDF030m	RDF descriptors	3D		F03[C−N]	2D frequency fingerprints	2D
	RDF055m	RDF descriptors	3D		F09[C−N]	2D frequency fingerprints	2D
	RDF020p	RDF descriptors	3D	THR	TI2	topological descriptors	2D
	RDF030p	RDF descriptors	3D		MATS5v	2D autocorrelations	2D
	Mor09u	3D-MoRSE descriptors	3D		EEig02x	edge adjacency indices	2D
	Mor04v	3D-MoRSE descriptors	3D		JGI7	topological charge indices	2D
	G3p	WHIM descriptors	3D		P2u	WHIM descriptors	3D
	H7m	GETAWAY descriptors	3D		E2p	WHIM descriptors	3D
	R6u	GETAWAY descriptors	3D		E3s	WHIM descriptors	3D
	R3m	GETAWAY descriptors	3D		H8m	GETAWAY descriptors	3D
	C-005	atom-centered fragments	1D		HATS6v	GETAWAY descriptors	3D
	H-047	atom-centered fragments	1D		nCq	functional group counts	1D
	N-072	atom-centered fragments	1D		nHDon	functional group counts	1D
	Hy	molecular properties	Others		H-053	atom-centered fragments	1D
	F01[O−S]	2D frequency fingerprints	2D		Hy	molecular properties	Others
	F07[N−F]	2D frequency fingerprints	2D		F08[C−S]	2D frequency fingerprints	2D
	RDF055m	RDF descriptors	3D		F08[N−O]	2D frequency fingerprints	2D

[a] Classification according to the descriptor family (or block as in DRAGON software). [b] Classification according to the dimensionality or complexity of molecular representation.

AUTHOR INFORMATION

Corresponding Author

*E-mail: oscarrb@uclv.edu.cu. Phone: 53 42 281515.

ACKNOWLEDGMENT

The first author (O.M.R.-B.) wishes to express his gratitude to Christof H. Schwab from Molecular Networks GmbH (Germany) for his kind support with the CORINA software. He is also grateful to Peter Willet from Sheffield University (England), János Podani from Eotvos University (Hungary), and Andreas Bender from Leiden University (Amsterdam) for their useful suggestions in the earliest stage of this research. He is also indebted to his colleagues and friends Noel Ferro from University of Hannover (Germany), Nelaine Mora-Diez from Thompson Rivers University (Canada), and Lourdes Casas-Cardoso from Cadiz University (Spain) for providing him with a lot of important bibliographical materials and Stephen J. Barigye (CAMD-BIR Unit, UCLV) for his revision of the English language. Last but not least, authors thank the anonymous reviewers for their exhaustive work throughout the manuscript resulting in a considerably improved scientific report. This research was partially supported by the UCLV and the VLIR-IUS Collaboration Programme. The Universidad Autónoma de Madrid—Universidad Central de Las Villas Scholarship Programme under the auspices of Caja Madrid, Spain also supported part of this research.

REFERENCES

(1) Jain, A. K.; Dubes, R. C. *Algorithms for Clustering Data*; Prentice-Hall: Englewood Cliffs, NJ, 1988.

(2) Jain, A. K.; Murty, M. N.; Flynn, P. J. Data clustering: A review. *ACM Comput. Surv.* **1999**, *31*, 264–323.

(3) (a) Downs, G. M.; Barnard, J. M. Clustering Methods and Their Uses in Computational Chemistry. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; John Wiley and Sons, Inc.: Hoboken, NJ, 2002; Vol. 18, pp 1−40; (b) Engels, M. F. M.; Gibbs, A. C.; Jaeger, E. P.; Verbinnen, D.; Lobanov, V. S.; Agrafiotis, D. K. A cluster-based strategy for assessing the overlap between large chemical libraries and its application to a recent acquisition. *J. Chem. Inf. Model.* **2006**, *46*, 2651–2660.

(4) (a) Adamson, G. W.; Bush, J. A. A method for the automatic classification of chemical structures. *Inf. Storage Retr.* **1973**, *9*, 561–568. (b) Adamson, G. W.; Bush, J. A. A comparison of the performance of some similarity and dissimilarity measures in the automatic classification of chemical structures. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 55–58.

(5) (a) Adamson, G. W.; Bawden, D. Comparison of hierarchical cluster analysis techniques for automatic classification of chemical structures. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 204–209. (b) Willett, P. A comparison of some hierarchical agglomerative clustering algorithms for structure—property correlation. *Anal. Chim. Acta* **1982**, *136*, 29–37. (c) Rubin, V.; Willett, P. A comparison of some hierarchal monothetic divisive clustering algorithms for structure-property correlation. *Anal. Chim. Acta* **1983**, *151*, 161–166. (d) Willett, P. Evaluation of relocation clustering algorithms for the automatic classification of chemical structures. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 29–33. (e) 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. Sci.* **1996**, *36*, 572–584. (f) Reynolds, C. H.; Druker, R.; Pfahler, L. B. Lead discovery using stochastic cluster analysis (SCA): A new method for clustering structurally similar compounds. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 305–312. (g) Holliday, J. D.; Rodgers, S. L.; Willett, P.; Chen, M.-Y.; Mahfouf, M.; Lawson, K.; Mullier, G. Clustering files of chemical structures using the fuzzy k-means clustering method. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 894–902. (h) Rodriguez, A.; Santos Tomas, M.; Perez, J. J.; Rubio-Martinez, J. Assessment of the performance of cluster analysis grouping using pharmacophores as molecular descriptors. *J. Mol. Struct.: THEOCHEM.* **2005**, *727*, 81–87.

(6) Willett, P. Searching techniques for databases of two- and three-dimensional chemical structures. *J. Med. Chem.* **2005**, *48*, 4183–4199.

(7) (a) Menard, P. R.; Lewis, R. A.; Mason, J. S. Rational screening set design and compound selection: Cascaded clustering. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 497–505. (b) Xu, J. A new approach to finding natural chemical structure classes. *J. Med. Chem.* **2002**, *45*, 5311–5320. (c) Luque Ruiz, I.; Cerruela García, G.; Gómez-Nieto, M. Á. Clustering chemical databases using adaptable projection cells and MCS similarity values. *J. Chem. Inf. Model.* **2005**, *45*, 1178–1194. (d) Stahl, M.; Mauser, H.; Tsui, M.; Taylor, N. R. A robust clustering method for chemical structures. *J. Med. Chem.* **2005**, *48*, 4358–4366. (e) Wilkens, S. J.; Janes, J.; Su, A. I. HierS: Hierarchical scaffold clustering using topological chemical graphs. *J. Med. Chem.* **2005**, *48*, 3182–3193. (f) Li, W. A fast clustering algorithm for analyzing highly similar compounds of very large libraries. *J. Chem. Inf. Model.* **2006**, *46*, 1919–1923. (g) Böcker, A.; Schneider, G.; Teckentrup, A. NIPALSTREE: A new hierarchical clustering approach for large compound libraries and its application to virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 2220–2229.

(8) Geppert, H.; Bajorath, J. Advances in 2D fingerprint similarity searching. *Expert Opin. Drug Discovery* **2010**, *5*, 529–542.

(9) (a) Haranczyk, M.; Holliday, J. Comparison of similarity coefficients for clustering and compound selection. *J. Chem. Inf. Model.* **2008**, *48*, 498–508. (b) Trepalin, S.; Yarkov, A. Hierarchical clustering of large databases and classification of antibiotics at high noise levels. *Algorithms* **2008**, *1*, 183–200.

(10) (a) Downs, G. M.; Willett, P.; Fisanick, W. Similarity searching and clustering of chemical—structure databases using molecular property data. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1094–1102. (b) Khalifa, A. A.; Haranczyk, M.; Holliday, J. Comparison of nonbinary similarity coefficients for similarity searching, clustering and compound selection. *J. Chem. Inf. Model.* **2009**, *49*, 1193–1201.

(11) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*, 2nd ed.; Wiley-VHC: Weinheim, Germany, 2009; Vols. I and II.

(12) Siegel, S.; Castellan, N. J. *Nonparametric Statistics for the Behavioral Sciences*; McGraw-Hill: New York, 1988.

(13) (a) Talavera, L. Dependency-based feature selection for clustering symbolic data. *Intell. Data Anal.* **2000**, *4*, 19–28. (b) Manoranjan, D.; Choi, K.; Scheuermann, P.; Huan, L. In *Feature Selection for Clustering: A Filter Solution*, Proceedings of the Second IEEE International Conference on Data Mining (ICDM'02), Maebashi City, Japan, December 9—12, 2002; IEEE Press: Maebashi City, Japan, 2002; pp 115—122; (c) Liu, T.; Liu, S.; Chen, Z.; Ma, W.-Y. In *An Evaluation on Feature Selection for Text Clustering*, Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, August 21—24, 2003; Fawcett, T., Mishra, N., Eds.; AAAI Press, Menlo Park, CA, 2003; pp 488—495; (d) Law, M. H. C.; Figueiredo, M. A. T.; Jain, A. K. Simultaneous feature selection and clustering using mixture models. *IEEE Trans. Pattern Anal.* **2004**, *26*, 1–13. (e) Raftery, A. E.; Dean, N. Variable selection for model-based clustering. *J. Am. Stat. Assoc.* **2008**, *101*, 168–178. (f) Yanjun, L. Text clustering with feature selection by using statistical data. *IEEE Trans. Knowl. Data* **2008**, *20*, 641–652.

(14) Böcker, A.; Derksen, S.; Schmidt, E.; Teckentrup, A.; Schneider, G. A hierarchical clustering approach for large compound libraries *J. Chem. Inf. Model.* **2005**, *45*, 807–815.

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

(16) (a) Biggs, J. B. The role of meta-learning in study process. *Br. J Educ. Psychol.* **1985**, *55*, 185–212. (b) de Souto, M. C. P.; Prudencio, R. B. C.; Soares, R. G. F.; de Araujo, D. S. A.; Costa, I. G.; Ludermir, T. B.; Schliep, A. In *Ranking and selecting Clustering Algorithms Using a Meta-Learning Approach*, Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN 2008), Hong Kong, China, June 1—8, 2008; Liu, D., Ed.; IEEE Press: Hong Kong, 2008; pp 3729—3735.

(17) Podani, J. New combinatorial clustering methods. *Vegetatio.* **1989**, *81*, 61–77.

(18) Sneath, P. H. A.; Sokal, R. R. *Numerical Taxonomy: The Principles and Practice of Numerical Classification*; W H Freeman & Co (Sd): San Francisco, 1973.

(19) Anderberg, M. R. *Cluster Analysis for Applications*. Wiley: New York, 1973.

(20) Lance, G. N.; Williams, W. T. A general theory of classificatory sorting strategies: 1. Hierarchical systems. *Comput. J.* **1967**, *9*, 373–380.

(21) (a) Jambu, M.; Lebeaux, M. O. Classification automatique pour l'analyse des données (1. Méthodes et algorithmes. 2. Logiciels). In *Dunod décision*; Dunod: Paris, 1978; pp 310—400; (b) Jambu, M.; Lebeaux, M. O. *Cluster Analysis and Data Analysis*; North-Holland: Amsterdam, 1983.

(22) Dubien, J. L.; Warde, W. D. A mathematical comparison of the members of an infinite family of agglomerative clustering algorithms. *Can. J. Stat.* **1979**, *7*, 29–38.

(23) Batagelj, V. Generalized Ward and Related Clustering Problems. In *Classification and Related Methods of Data Analysis*; Bock, H. H., Ed.; North-Holland: Amsterdam, 1988; pp 67—74.

(24) Hubálek, Z. Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. *Biol. Rev.* **1982**, *57*, 669–689.

(25) Murtagh, F. A survey of recent advances in hierarchical clustering algorithms. *Comput. J.* **1983**, *26*, 354–359.

(26) Maldonado, A. G.; Doucet, J. P.; Petitjean, M.; Fan, B.-T. Molecular similarity and diversity in chemoinformatics: From theory to applications. *Mol. Diversity* **2006**, *10*, 39–79.

(27) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7*, 903–911

(28) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A comparison of methods for modeling quantitative structure—activity relationships. *J. Med. Chem.* **2004**, *47*, 5541–5554.

(29) (a) Bruce, C. L.; Melville, J. L.; Pickett, S. D.; Hirst, J. D. Contemporary QSAR classifiers compared. *J. Chem. Inf. Model.* **2007**, *47*, 219–227. (b) Culp, M.; Johnson, K.; Michailidis, G. The ensemble bridge algorithm: A new modeling tool for drug discovery problems. *J. Chem. Inf. Model.* **2010**, *50*, 309–316. (c) Johansson, U.; Löfström, T.; Norinder, U. In *Evaluating Ensembles on QSAR Classification*, Proceedings of the 3rd Skövde Workshop on Information Fusion Topics 2009 (SWIFT 2009), Skövde, Sweden; Johansson, R., van Laere, J., Mellin, J.,

Eds.; Univeristy of Skövde: Skövde, Sweden, 2009; pp 49−54. (d) Sönströd, C.; Johansson, U.; Norinder, U. In *Generating Comprehensible QSAR Models*, Proceedings of the 3rd Skövde Workshop on Information Fusion Topics 2009 (SWIFT 2009), Skövde, Sweden; Johansson, R., van Laere, J., Mellin, J., Eds.; University of Skövde: Skövde, Sweden, 2009; pp 44−48.

(30) Johnson, M. A. A review and examination of mathematical spaces underlying molecular similarity analysis. *J. Math. Chem.* **1989**, *3*, 117–145.

(31) (a) Maggiora, G. M.; Shanmugasundaram, V. Molecular Similarity Measures. In *Chemoinformatics*; Bajorath, J., Ed.; Humana Press: New York, 2004; Vol. 275, pp 1−50; (b) Agrafiotis, D. K.; Bandyopadhyay, D.; Wegner, J. K.; van Vlijmen, H. Recent advances in chemoinformatics. *J. Chem. Inf. Model.* **2007**, *47*, 1279–1293.

(32) Bender, A.; Glen, R. C. Molecular similarity: A key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204–3218.

(33) Janecek, A.; Gansterer, W.; Demel, M.; Ecker, G. In *On the Relationship between Feature Selection and Classification Accuracy*, Proceedings of the Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery (FSDM 2008), Antwerp, Belgium, September 15, 2008; Saeys, Y., Liu, H.; Inza, I., Wehenkel, L., Van de Peer, Y., Eds.; JMLR: Workshop and Conference Proceedings: Antwerp, Belgium, 2008; pp 90−105.

(34) Steinbach, M.; Ertöz, L.; Kumar, V. The Challenges of Clustering High Dimensional Data. In *New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition*; Wille, L. T., Ed.; Springer-Verlag: Berlin/New York, 2000; pp 273−307.

(35) John, G. H.; Kohavi, R.; Pfleger, K. In *Irrelevant Features and the Subset Selection Problem*, Proceedings of the Eleventh International Conference on Machine Learning (ICML), Rutgers University, New Brunswick, NJ, USA; Cohen, W. W.; Hirsh, H., Eds.; Morgan Kaufman: NJ, 1994; pp 121−129.

(36) Watanabe, S. *Knowing and Guessing: A Quantitative Study of Inference and information*; John Wiley & Sons, Inc: New York, 1969.

(37) Böcker, A.; Schneider, G.; Teckentrup, A. Status of HTS data mining approaches. *QSAR Comb. Sci.* **2004**, *23*, 207–213.

(38) (a) *JChem for Excel*, 5.3.8 (166); Budapest, Hungary, 2010. (b) JChem for Excel is a Microsoft Excel integrated tool enabling scientists to manage and analyze chemical structures and their data. The software is available from ChemAxon Kft. at http://www.chemaxon.com (accessed July 27, 2011).

(39) (a) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of automatic three-dimensional model buildersusing 639 X-ray structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000–1008. (b) The 3D structure generator CORINA is available from Molecular Networks GmbH at http://www.molecular-networks.com (accessed July 27, 2011).

(40) (a) *DRAGON for Windows*, 5.5; Milano, Italy, 2007. (b) The software for molecular descriptors calculations DRAGON is available from Talete srl at http://www.talete.mi.it (accessed July 27, 2011).

(41) (a) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA data mining software: An update. *SIGKDD Explor. Newsl.* **2009**, *11*, 10–18. (b) Weka is a collection of machine learning algorithms for data mining tasks. The software Weka v. 3-6-4 is available from the Machine Learning Group at University of Waikato at http://www.cs.waikato.ac.nz/ml/weka/ (accessed July 27, 2011).

(42) Hall, M. A. Correlation-Based Feature Subset Selection for Machine Learning. PhD. Thesis, The University of Waikato, Hamilton, New Zealand, 1998.

(43) (a) Podani, J. *SYN-TAX2000*; Scientia Publishing: Budapest, Hungary, 2001. (b) The SYN-TAX program package is designed for multivariate data analysis in SYNbiology (or Ecology) and TAXonomy (or Systematics). It is available from request to Professor János Podani at http://ramet.elte.hu/∼podani/subindex.html (accessed July 27, 2011).

(44) Podani, J. A method for generating consensus partitions and its application to community classification. *Coenoses* **1989**, *4*, 1–10.

(45) Podani, J. Explanatory Variables in Classifications and the Detection of the Optimum Number of Clusters. In *Data Science, Classification and Related Methods*; Hayashi, C., Ohsumi, N., Yajima, K., Tanaka, Y., Bock, H. H., Eds.; Springer: Tokyon, 1998; pp 125−132.

(46) Stein, B.; Meyer zu Eissen, S.; Wißbrock, F. In *On Cluster Validity and the Information Need Of users*, Proceedings of the 3rd IASTED International Conference on Artificial Intelligence and Applications (AIA 03), Benalmádena, Spain; Hanza, M. H., Ed.; ACTA Press: Benalmádena, Spain, 2003; pp 216−221.

(47) Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A. F.; Nielsen, H. Assesing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics* **2000**, *16*, 412–424.

(48) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046–1053.

(49) Wolpert, D. H. In *The Supervised Learning No-Free-Lunch Theorems*, Presented at the 6th Online World Conference on Soft Computing in Industrial Applications (WSC6). [online], September 10−24, 2001. http://ti.arc.nasa.gov/profile/dhw/statistical/ (accessed July 27, 2011).

(50) Kruskal, W. H.; Wallis, W. A. Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* **1952**, *47*, 583–621.

(51) Conover, W. J.; Iman, R. L. Rank transformations as a bridge between parametric and nonparametric statistics. *Am. Stat.* **1981**, *35*, 124–129.

(52) (a) Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30. (b) García, S.; Herrera, F. An extension on "Statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *J. Mach. Learn. Res.* **2008**, *9*, 2677–2694.

(53) (a) Milligan, G. W. Ultrametric hierarchical clustering algorithms. *Psychometrika* **1979**, *44*, 343–346. (b) Batagelj, V. Note on ultrametric clustering algorithms. *Psychometrika* **1981**, *46*, 351–352.

(54) Diday, E. Inversions en classification hiérarchique: Application á la construction adaptive d'indices d'agrégation. *Rev. Stat. Appl.* **1983**, *31*, 45–62.

(55) Fawcett, T. An introduction to ROC analysis. *Pattern Recogn. Lett.* **2006**, *27*, 861–874.

(56) (a) Ivanciuc, O. Applications of Support Vector Machines in Chemistry. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Cundari, T. R., Eds.; Wiley-VCH: Weinheim, Germany, 2007; Vol. 23, pp 291−400. (b) Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: Foundations, limitations and novel approaches. *Drug Discovery Today* **2007**, *12*, 225–233.