JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

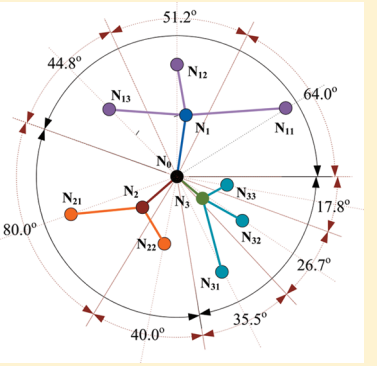
pubs.acs.org/jcim

Analysis and Study of Molecule Data Sets Using Snowflake Diagrams of Weighted Maximum Common Subgraph Trees

Gonzalo Cerruela García, Irene Luque Ruiz,* and Miguel Ángel Gómez-Nieto

Department of Computing and Numerical Analysis, University of Córdoba, Campus de Rabanales, Albert Einstein Building, E-14071 Córdoba, Spain

ABSTRACT: Isomorphism measures based on the maximum common subgraph (MCS) calculation are widely used in computational chemistry for classifying, screening, and predicting properties and biological activity within chemical databases. The development of a weighted hierarchical structure based on the MCS is described in this paper. Furthermore, a 2D representation model is proposed as the proper tool for the study and preliminary analysis of molecule data sets. The development process of the weighted MCS tree is open to the use of different approaches. By taking into account different molecular descriptors, similarity and distance measures in the weighted MCS tree, the relationships between the molecular property or the activity, and the variables considered for the building and display of the weighted MCS tree can be observed. Besides that, the representation model based on snowflake diagrams allows to display of those relationships as well as shows any existing degeneration, in order to detect any possible outlier that could be obtained during the development of predictive models and to extract new variables that can be used in the building of quantitative structure−activity relationship models.



1. INTRODUCTION

The use of molecule data sets representation techniques is a basic and previous task for the further development of other activities aimed at different aspects of the investigation or development in computational chemistry, chemometric, quantitative structure−property relationship/quantitative structure−activity relationship (QSPR/QSAR), clustering and screening of chemical databases, and so on.[1−6]

Traditionally, the representation techniques used have been based on two steps: (a) The first is to obtain a space for the representation of the molecule data set, and (b) the second is to apply some transformations to this representational space through various functions in order to obtain a point for each element of the data set in the Euclidean space used by the representation technique.

The main aim of graphic representations of molecule data sets is to show a graphical view of the characteristics of the data set, so conclusions can be obtained, and decisions can be made without costly computational calculations and/or choosing/synthesizing other chemical compounds.

Among the most generalized methods for molecule data set representations are those techniques used in database clustering and screening, and lately some techniques focused on finding cliffs that may provoke notorious deviations on the development of predictive models for biological properties and activities. Within them, the hierarchical classification methods, K-means, principal components analysis, etc.,[7−9] have been and will be traditionally used as the initial step in the data set study. These techniques use a representational space of the molecules data set based, mainly, on variables or descriptors extracted from the molecular structure, an isomorphic representation of the molecule (molecular graph, reduced graph, etc.), fingerprints, etc.[10−14]

Recently, some investigations have been carried out with the result of revealing that a convenient way for representing the structure of the molecules of the data set is through an array containing the common and uncommon fragments of the elements of the data set. Thus, making use of matching and fragmenting algorithms, molecules data sets are displayed in a hierarchical structure (spanning tree) where the root node corresponds to the most common fragment of the elements of the data set, branch nodes represent the elements of the data set, and child nodes represent common fragments to subsets of the elements of the data set.

Using the spanning tree built as the starting point, different representation models have been proposed with the aim of performing a quick and efficient analysis of the features of the data set in researches related to cliff and outlier detection in the development of predictive models (SAR)[15−18] as well as in the screening and categorization of large chemical databases. Generally, those diagrams have in common that the root node of the spanning tree is displayed in the origin of the coordinates, and the other nodes of the tree are displayed in a circular space around the center, with each node of the spanning tree being assigned a point in the space.[19−21] This point has been calculated based on different considerations (sometimes on more than only one):

- Features or descriptors drawn from the fragment represented by the node.

Received: December 8, 2010
Published: April 24, 2011

- Variables obtained based on the existing relationships among the nodes of the spanning tree.
- Similarity, possibility, etc. measures among the node of the spanning tree.
- Topological measures derived from a given representation model.

It can be observed that representations generated are very similar in all cases, distributing the nodes of the spanning tree around a center (root node) and presenting a circular distribution in branches that expand themselves over their own central node. These kinds of representations, some time called star-like diagrams, have been widely used in several fields of computer sciences (data mining, logic, knowledge semantic representation, networks, etc.), under different names, such as snowflake diagrams, in many areas of computer science.[22−27]

In this paper we propose the use of snowflake diagrams for the representation of weighted maximum common subgraph trees (WMCST). WMCST are balanced spanning trees where the root node represents the maximum common structure (MCS) common to the whole data set, the branch nodes are the molecules of the data set, and the internal nodes are the fragments that contain the MCS common to subsets of molecules of the data set. In the WMCST both the nodes and the arcs are weighted, this weight being a feature or property of the node (fragment) or of the arc (relationship between fragments or the cost of getting from one fragment to another).

In this paper we describe the characteristics and the method for the generation of the WMCST snowflake diagram, outlining the advantages that it provides for the quick and efficient analysis of families of molecules, in activities of the initial analysis, diminishing costly QSAR subsequent studies, clustering, etc. In further papers we will highlight the advantages of this type of representation against other techniques, its use in cliff detection, and the possibility it offers to generate new representational spaces of the data set for the development of predictive models.

The paper has been structured in the following way: In Section 2 the algorithm and the procedures used for the data set fragmentation and the development of a MCS spanning tree are described. In Section 3 the basis for the development of the WMCST is introduced, describing the representational space for the snowflake diagram, the node distribution within the diagram, and how the method is open to different models of weighting. In Section 4 the features of these diagrams are described, specifically for the example data set, describing the impact of the parameters of the snowflake diagram and analyzing the data set behavior and biological activity in relation to these parameters. Finally, advantages of these kinds of diagrams for the initial analysis of molecules data sets with the objective of developing classification processes and building predictive models are discussed.

2. MATERIALS AND METHOD

In several applications in computational chemistry it is quite common to simplify the isomorphism of molecular graphs to just finding the MCS with the aim of obtaining similarity measures among molecules. Similarity measures obtained are widely used in chemical application, such as database searching, biological activity prediction, molecular spectrum interpretation, etc.[28−31]

The problem of getting the MCS from two molecular graphs is a high computational cost problem with different efficient solutions proposed by different researchers. In this problem the computational cost increases with the measure of the data set under study. Recently[32] a new hierarchical classification algorithm based on the MCS tree (MCST) has been proposed. In this case the number of operations needed is widely reduced, and the outliers are classified as singletons, avoiding the fragmentation of common structures of bigger size.

The algorithm starts by classifying all the structures on the data set in branches (bottom) of the hierarchy. The next level contains the MCSs as clusters of the initial molecules, all the molecules that share a common structure are assigned to the same cluster. At the root of the hierarchy there is one (or many) MCS common to the whole data set, creating disjoint clusters (one molecule belongs to just one unique cluster).

One implementation of this algorithm,[33] uses different heuristic techniques for minimizing the computational cost in the development of the MCST: (a) use of the similarity matrix obtained from the data set fingerprints for predefining a similarity threshold as main condition for the calculation of the MCS, (b) no consideration of MCSs that could contain a number of atoms under a preset value, and (c) two classification models, one approximate and one exact. In this paper the libMCS[33] implementation of the algorithm has been used for the development of the MCST.

Once the MCST has been obtained, our work consists of building the weighted maximum common subgraph tree (WMCST), which will be described in the following section, creating a hierarchical structure that stores information related to:

- The substructures uncommon between each node of the hierarchy. So for each level all the child nodes of the same parent node share the same MCS, and each node will store only the uncommon structures with its parent.
- Structural similarity coefficient of each node with regards to the root or parent node. These values will be used for measuring the distance or weigh of the arcs in the WMCST.
- Value of the molecular descriptor considered in each node of the tree and of each uncommon substructure. These values will be used for measuring the weight of the nodes and the distance of the arcs in the WMCST.
- Value of the property under analysis for each node of the tree. These values will be used for measuring the weight of the nodes and the distance of the arcs in the WMCST. For those nodes of the WMCST that do not belong to molecules on the data set, those values will be calculated based on the values of the property of the molecules assigned to the same cluster.

The algorithm for building the WMCST has been developed in Java[34] and integrated into the CoChiSE[35,36] tool for the calculation of the molecular descriptors, similarity measures, and extraction of nonisomorphic substructures among the nodes of the same cluster.

3. THEORETICAL BASIS

3.1. Building the WMCST. A WMCST is a hierarchical structure where the leaf nodes represent the elements of a molecule data set, the child nodes represent common substructures to subsets of the data set, and the root node (unique) has the MCS common to all the elements on the data set. In this structure, each node of the tree has a weight assigned $W(N_i)$, and the relationships between all the nodes (the arcs) have a cost or distance assigned to them $d(N_i, N_j)$.

In the WMCST the label or the weight assigned to a node $W(N_i)$ is a property defined solely by the features of the

substructure or molecule represented by that node, while the label or the distance assigned to the arc that ties the nodes N_i and N_j is determined by the structure characteristics of both nodes N_i and N_j. So, while $W(N_i)$ can be obtained from the $f(MS_i)$ function, MS_i being the molecular structure represented by N_i, $d(N_i, N_j)$ must be obtained from a $g(MS_i, MS_j)$ function, which means the cost of getting or reaching MS_j from MS_i. Thus, examples of the $f()$ function may be: Wiener, Randic, Szegel, or any other molecular index, value of the property in study, etc., and examples of the $g()$ function may be: the value of any molecular descriptors over the nonisomorphic fragment obtained in the matching of parent and child node, the difference of property in study, the difference between the weight between parent and child node, the energy necessary to obtain child node from parent node, etc.

3.1.1. Calculating the Weight of the Nodes. In the WMCST the weight assigned to the N_i node represents a structural feature of the MS_i substructure assigned to that node. This weight is calculated through a $f()$ function whose objective is to get the value of a descriptor or a variable that allows us to display the MS_i substructure.

During recent decades, hundreds of molecular descriptors have been proposed[37] which could be calculated from different structural representations of the molecules or molecules' fragments. The calculation of many molecular descriptors, mainly topological, is based on:

- Representing the structure of the molecules or chemical compounds through a graph (molecular graph), where the nodes are the atoms and the edges are the bounds. This graph may, or may not, be colored. In that case information about the characteristics of the nodes and the arcs is kept. Colored molecular graphs can store different characteristics about the nodes and arcs (i.e., kind of atom and bound, distance of the bound, atomic weight, electro negativity, etc.).
- Changing the abstract representation of the molecular graph into a matrix representation so it can be easily and efficiently processed by a computer. Depending on whether the molecular graph is colored and the kind of feature or property stored in the nodes and arcs, different matrix representations have been proposed (adjacencies, connections, electronegativity, etc., matrices).
- And, finally, obtaining an invariant or descriptor from the matrix that represents the molecular graph. In this process, algorithms may generate other matrix representations from the initial one. Generally, in the process a sole value or invariant is obtained, although different molecular descriptors generate different values that are defined as molecular invariants.

The existence of hundreds, or thousands, of molecular descriptors is based on the nature of the calculation process and on the molecular feature that would be displayed thanks to the invariant obtained. As different features can be considered in the colored molecular graph that represents the molecule, the molecular descriptor calculated has as its main objective to obtain a relationship with the shape: $p = h(D_M)$, being p a physic−chemic property or a biological activity, and D_M the value of the molecular descriptor obtained. Being this the foundation of the analysis performed in QSPR/QSAR, where for the investigation of a property p, the use of the molecular descriptors D_M have been proposed regarding the characteristics associated to the arcs and nodes of the colored molecular graph and the algorithm used for the invariant calculation.

In our model the $f()$ function can be any algorithm that generates an invariant or molecular descriptor (D_M) of the substructure (MS_i) represented by the N_i node. For this calculation the aforementioned transformation process takes place, obtaining, for each data set under analysis, the most appropriate weight (descriptor value) regarding the property studied.

3.1.2. Calculating the Weight of the Edges. In the WMCST the weight of the arcs represents the distance or cost of getting or obtaining the node N_j from the node N_i. From a formal point of view, the distance between two nodes N_i and N_j can be defined as any $g(N_i, N_j)$ function that fulfills:

- The distance has a positive or zero value: $g(N_i, N_j) \geq 0$, $\forall N_i, N_j$.
- The distance is symmetric: $g(N_i, N_j) = g(N_j, N_i)$.
- The distance is not cumulative: $g(N_i, N_k) \leq g(N_i, N_j) + g(N_j, N_k), \forall N_i, N_j, N_k$. In case the latter is not fulfilled, the distance is called pseudodistance or pseudometric.

Different criteria may be taken into account in the WMCST for the calculation of the distance between two nodes. A criterion directly inferred from the nature of the WMCST tree would be based on the calculation of the distance as 1 minus the value of the similarity (S) between the substructures represented by the nodes N_i and N_j as shown in eq 1:

$$d(N_i, N_j) = 1 - S(MS_i, MS_j) \quad (1)$$

This calculation can be done making use of different similarity models and indexes, always obtaining a measure within the [0−1] range, i.e.:

- Structural similarity measures could be used when accounting the elements (nodes and edges) of the molecular graphs of the substructures assigned to each node and of the common elements (nodes and edges).
- Similarity measures based on fingerprints could be used. For this purpose the fingerprints corresponding to the substructures MS_i and MS_j might be built first.
- Measures based on approximate similarity[38,39] could also be used. These metrics take into account both structural similarity and fingerprints measures as well as the use of molecular descriptors for the fine-tuning of the calculation. It has been demonstrated that by these measures, a quite proper similarity measure for the development of QSPR/QSAR models could be obtained.

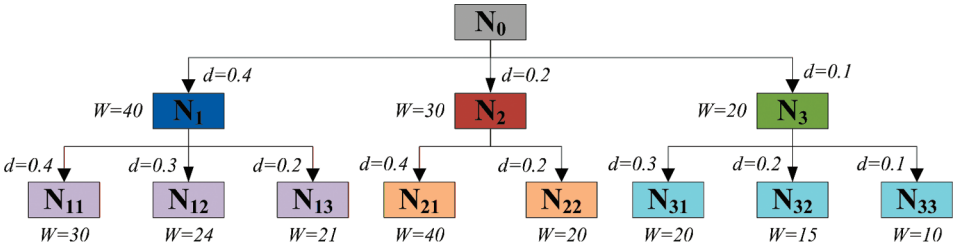
If we take into account the use of structural similarity measures, as per the node N_i (parent) has always a size minor or equal to the node N_j, as per the MS_i substructure is a fragment of the MS_j substructure (or it is the same), the similarity value obtained would be based on the difference in size between both structures. Thus, the size of the noncommon fragment to both substructures MS_i and MS_j would determine the similarity value. So, the more the difference in size between MS_i and MS_j is, the more the similarity diminishes, and therefore, the distance between N_i and N_j increases.

Apart from that, bearing in mind that the WMCST is a hierarchical structure where the parent node of a specific branch represents the MCS common to all the child nodes of the branch, the distance between a parent and a child should be given by

$$d(N_i, N_j) = g(\mathrm{NIF}[MS_i, MS_j]) \quad (2)$$

where $\mathrm{NIF}[MS_i, MS_j]$ is the nonisomorphic substructure or fragment between the MS_i and MS_j substructures, represented by

Table 1. An Example of MCST and Information Corresponding with the WMCST Diagram of Figure *1*



node	weight	distance	circular arc	initial angle	final angle	point
N_0	n/a	0.0	360.0	0.0	360.0	(0.00, 0.00)
N_1	40	0.4	(40/90) × 360 = 160.0	0.0	160.0	(0.07, 0.39)
N_2	30	0.2	(30/90) × 360 = 120.0	160.0	280.0	(−0.15, −0.13)
N_3	20	0.1	(20/90) × 360 = 80.0	280.0	360.0	(0.08, −0.06)
N_{11}	30	0.4	(30/75) × 160 = 64.0	0.0	64.0	(0.68, 0.42)
N_{12}	24	0.3	(24/75) × 160 = 51.2	64.0	115.2	(0.00, 0.70)
N_{13}	21	0.2	(21/75) × 160 = 44.8	115.2	160.0	(−0.44, −0.40)
N_{21}	40	0.4	(40/60) × 120 = 80.0	160.0	240.0	(−0.56, −0.21)
N_{22}	20	0.2	(20/60) × 120 = 40.0	240.0	280.0	(−0.07, −0.39)
N_{31}	20	0.3	(20/45) × 80 = 35.5	280.0	315.5	(0.19, −0.35)
N_{32}	15	0.2	(15/45) × 80 = 26.7	315.5	342.2	(0.26, −0.16)
N_{33}	10	0.1	(10/45) × 80 = 17.8	342.2	360.0	(0.20, −0.03)

the nodes N_i and N_j, and $g()$ is a function in charge of obtaining a descriptor or invariant from NIF[MS_i, MS_j].

The concept of NIF has been previously used in computational chemistry for the proposal of classification and screening models for chemical databases and for the proposal of models for the prediction of physic−chemical properties and biological activity.[39,41] Its consideration allows the use of molecular descriptors for the calculation of invariants that could be later used for different applications, as in the correction of the structural similarity value and the obtainment of an approximate similarity measure that permits, bearing in mind structural and other aspects, defined by the molecular descriptor used in its calculation.

The use of the eq 2, allows us to get distance measures greater than 1, regarding the descriptor used for the $g()$ function; although the nature of the descriptor used will determine whether the distance is a metric or a pseudometric.

Furthermore, the distance calculation can be done taking into account different criteria:

- All the distances are referred to the root node of the hierarchical structure; this means, to the substructure for what MCS is common to all the molecules of the data set. Thus, the distance between N_i and N_j is equal to

$$d(N_i, N_j) = \sqrt{[g(N_0, N_j)]^2 - [g(N_0, N_i)]^2} \quad (3)$$

  where: N_0 represents the root node of the tree, herewith, the MCS common to the data set.
- The distance calculation is made by levels, only taking into account the parent and child nodes.

The values obtained in each case would be equal or not, regarding if the criteria used in the distance calculation (equation) and the characteristics of the $g()$ function.

3.1.3. The Weighted MCS Tree. Once the criteria used for the calculation of the weights of each node and the distance between nodes or weight of the arcs is selected, it is possible to build up the WMCST and its representation through a matrix structure that is easily processed. A square matrix, with the same number of rows and columns as the number of nodes of the MCST tree, where the main diagonal stores the weight of each node, and the (i, j) elements store the weight of the arcs, or zero in the case that node j is not a child node of the i node.

3.2. Building the Snowflake Diagram for the WMCST. A snowflake diagram is a graphic representation of a set of data that mimics the structure of crystallized water. This type of representation has been used in several areas, with the objective of having a flat and circular representation of the distribution of a set of data with parent−child relationships (hierarchy).

The snowflake diagram for the WMCST is based on the following principles:

(a) The WMCST is displayed over an *XY* plane, where the root node of the WMCST tree is assigned to the center of coordinates.

(b) Nodes of the first level are represented in a circle whose center is the center of coordinates in the following way:

(i) An arc of the circle is assigned to each node. This arc is proportional to the weight of the node compared to the total weight of the nodes from the first level ($l = 1$), that means:

$$\text{arc}(N_{l,j}) = \frac{W(N_{l,j})}{\sum_{i=1}^{i=k} W(N_{l,i})} \times 360° \quad (4)$$

(ii) The assignment is made in descending order regarding the value of the weight of each node, starting from 0°, so

dx.doi.org/10.1021/ci100484z |*J. Chem. Inf. Model.* 2011, 51, 1216–1232

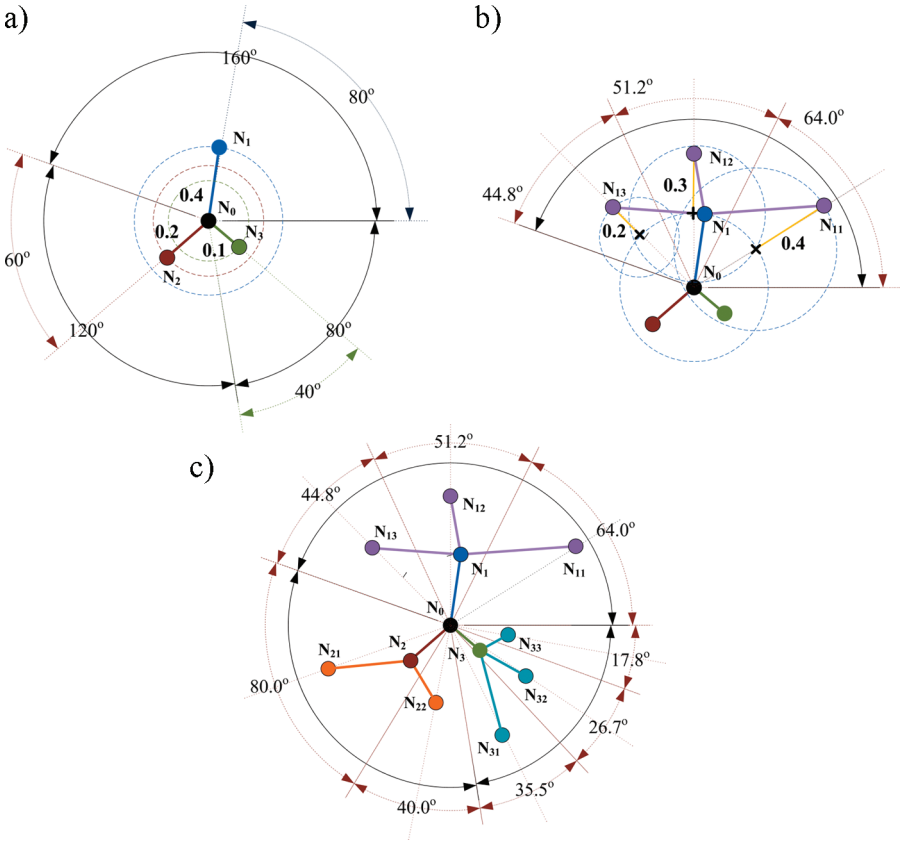


Figure 1. Representation method for the snowflake diagram.

each node has assigned an initial angle and a value for the arc of the circle that would contain its child nodes.

(c) The place assigned to each node of the first level is calculated in the following way:

(i) A circle with radius equal to the distance between the root node (center of the circle) and each node of the first level is drawn. This node could be assigned to any point in the circumference.

(ii) The bisectrix of the quarter assigned to each node of the first level is obtained, as well as the intersection point of the bisectrix with the circumference's arc that belongs to the quarter assigned to each node, this being the point assigned to the child node.

(d) After that, the root node and each one of the nodes of the first level are linked.

Finally, the points assigned to the nodes of the first level can be obtained from solving the following equations system (eqs 5 and 6) for each one of the points:

$$d_i^2(\mathrm{N}_0, \mathrm{N}_i) = x_i^2 + y_i^2 \tag{5}$$

$$y_i = tag\left(\sum_{j=1}^{j=i-1} [\mathrm{arc}(\mathrm{N}_j)] + \frac{\mathrm{arc}(\mathrm{N}_i)}{2}\right) x \tag{6}$$

For solving the points assigned to the nodes of the following levels, the same process previously described must be executed, taking into account that the nodes of level k have a circumference arc assigned that is related and included within the circumference arc assigned to the parent node of the level $k - 1$.

Table 1 shows a WMCST example built in order to clarify the explanation of building snowflake diagrams. In the WMCST of Table 1, we have considered 8 molecules (named N_{11}−N_{13}, N_{21} and N_{22}, and N_{31}−N_{33}). Molecules N_{11}−N_{13} have a common MCS (structure N_1), molecules N_{21} and N_{22} have the structure N_2 as common MCS, and molecules N_{31}−N_{33} have the structure N_3 as common MCS. Thus, structures N_1−N_3 compose the first level of the MCST. The root node of MCST is the node N_0, that is the MCS common to the entire data set.

The weight of the nodes and the distance between nodes of MCST example are calculated, and the WMCST is generated. Table 1 shows the WMCST and the weight and distance information. This WMCST example is used for explaining the simplified building of a snowflake diagram represented in Figure 1.

In the example of Figure 1, the root node (N_0) has three child nodes with weights of 40, 30, and 20 and arc distances of 0.4, 0.2, and 0.1 respectively. The node N_1 has also 3 child nodes with weights of 30, 24, and 21 and distances of 0.4, 0.3, and 0.2, respectively. The node N_2 has two child nodes of weights of 40 and 20 and distances of 0.4 and 0.2, respectively, and finally, the node N_3 has also 3 child nodes with weights of 20, 15, 10 and distances of 0.3, 0.2, and 0.1, respectively.

In the first step, the root node (N_0) is assigned to the center of the circumference, and each child node (N_1−N_3) is assigned an arc regarding its weight, so the values included in Table 1 are displayed and shown in Figure 1a.

Then, and for each child node, the circumference of radius $d(N_0, N_i)$ is obtained, for example, a circumference of radius 0.4 for the node N_1. This N_1 could be displayed at any point of the circumference. The point chosen is the point where the bisectrix

dx.doi.org/10.1021/ci100484z |J. Chem. Inf. Model. 2011, 51, 1216–1232

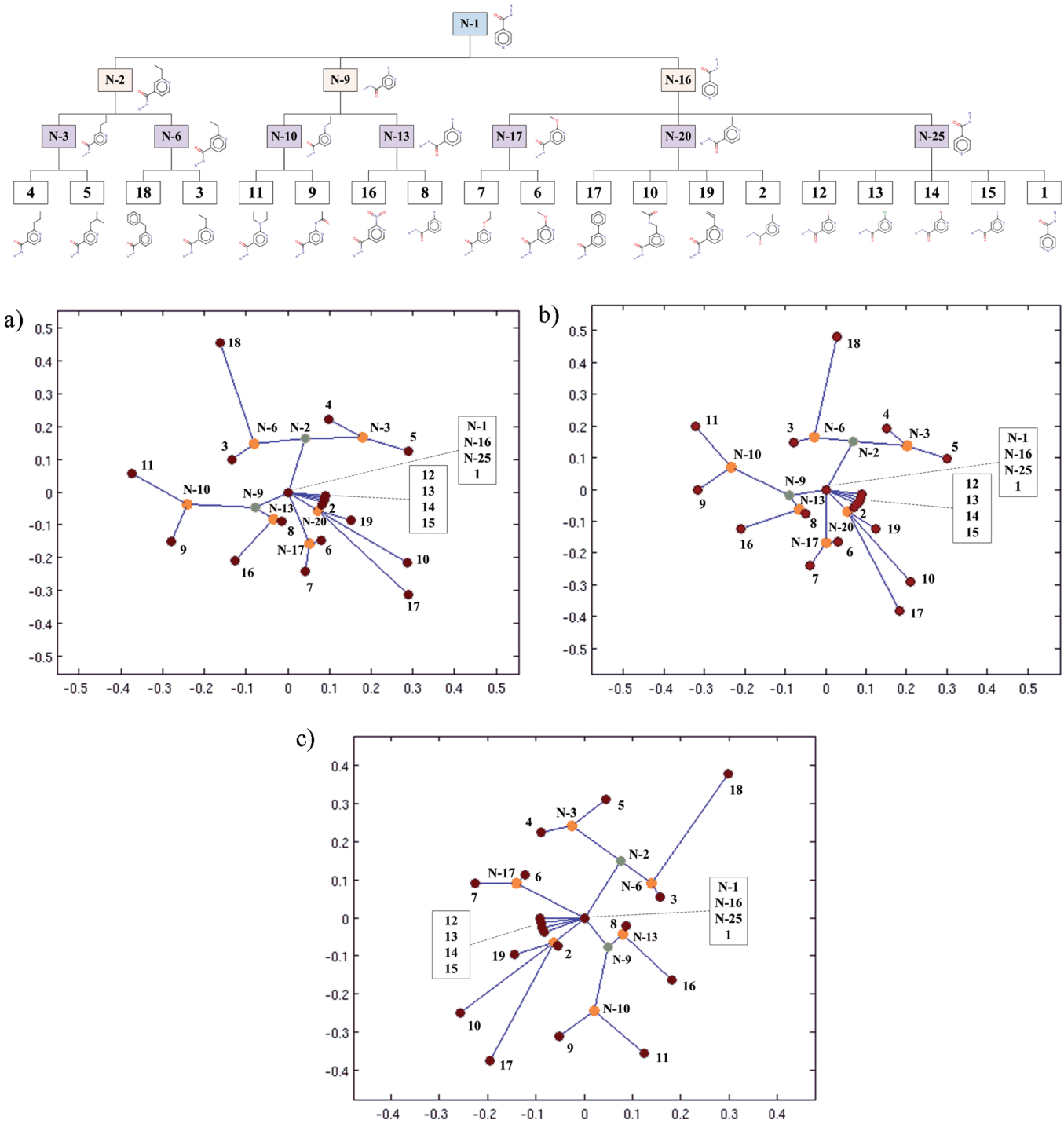


Figure 2. Behavior of the weight assigned to the nodes in the WMCST snowflake diagram. MCST (top), Wiener, Randic, and Balaban descriptors based on snowflake diagrams.

intersects the quarter assigned to the node N_1. And so, the process is repeated for the other nodes of the first level, as in the diagram of Figure 1a.

For calculating the points where the nodes of the second level are displayed, the same process is performed. For example, for the child nodes of the node N_1 (nodes N_{11}−N_{13}) the following steps are performed (see Figure 1b):

(a) Each node is assigned a quarter of the circumference regarding its weight and embedded within the circumference's quarter assigned to the parent node (see Table 1). For example, the node N_1 has a circumference quarter within 0° and 160°. So the node N_{11} is assigned a quarter within 0° and 64°, the node N_{12} is assigned a quarter within 64° and 115.2°, and the node N_{13} the quarter within 115.2° and 160°.

(b) For each child node the circumference with radius equal to the distance to the parent node and with center equal to the intersection of the bisectrix of the quarter assigned

dx.doi.org/10.1021/ci100484z |J. Chem. Inf. Model. 2011, 51, 1216–1232

Table 2. Experimental Results for Hydrazide Data Set with Different Molecular Descriptors Weighting the Nodes of the WMCST[a]

			Wiener						Randic						Balaban					
L	ID	F	*I*	*D*	*X*	*Y*	α_i	α_f	*I*	*D*	*X*	*Y*	α_i	α_f	*I*	*D*	*X*	*Y*	α_i	α_f
0	N-0	N-0	0.000	0.000	0.000	0.000	0.000	360.000	0.000	0.000	0.000	0.000	0.000	360.000	0.000	0.000	0.000	0.000	0.000	360.000
1	N-1	N-0	121.000	0.000	0.000	0.000	0.000	360.000	4.843	0.000	0.000	0.000	0.000	360.000	2.240	0.000	0.000	0.000	0.000	360.000
2	N-16	N-1	121.000	0.000	0.000	0.000	269.061	360.000	4.843	0.000	0.000	0.000	250.035	360.000	2.240	0.000	0.000	0.000	125.976	246.315
2	N-2	N-1	202.000	0.167	0.041	0.162	0.000	151.816	5.774	0.167	0.069	0.152	0.000	131.126	2.344	0.167	0.076	0.000	0.000	125.976
2	N-9	N-1	156.000	0.091	−0.078	−0.046	151.816	269.061	5.236	0.091	−0.089	−0.017	131.126	250.035	2.116	0.091	0.050	0.000	246.315	360.000
3	N-10	N-9	260.000	0.154	−0.242	−0.036	151.816	225.094	6.274	0.154	−0.235	0.069	131.126	195.942	2.143	0.154	0.021	0.000	246.315	303.518
3	N-13	N-9	156.000	0.000	−0.035	−0.084	225.094	269.061	5.236	0.000	−0.067	−0.062	195.942	250.035	2.116	0.000	0.080	0.000	303.518	360.000
3	N-17	N-16	202.000	0.167	0.052	−0.158	269.061	307.411	5.774	0.167	0.000	−0.167	250.035	290.088	2.344	0.167	−0.140	0.000	125.976	168.086
3	N-20	N-16	156.000	0.091	0.072	−0.056	307.411	337.028	5.236	0.091	0.056	−0.071	290.088	326.410	2.116	0.091	−0.062	0.000	208.312	246.315
3	N-25	N-16	121.000	0.000	0.000	0.000	337.028	360.000	4.843	0.000	0.000	0.000	326.410	360.000	2.240	0.000	0.000	0.000	168.086	208.312
3	N-3	N-2	260.000	0.077	0.179	0.165	0.000	85.438	6.274	0.077	0.202	0.137	0.000	68.284	2.143	0.077	−0.025	0.000	65.820	125.976
3	N-6	N-2	202.000	0.000	−0.080	0.146	85.438	151.816	5.774	0.000	−0.028	0.164	68.284	131.126	2.344	0.000	0.140	0.000	0.000	65.820
4	**1**	N-25	121.000	0.000	0.000	0.000	356.269	360.000	4.843	0.000	0.000	0.000	353.692	360.000	2.240	0.000	0.000	0.000	168.086	176.504
4	**2**	N-20	156.000	0.000	0.083	−0.038	333.209	337.028	5.236	0.000	0.072	−0.055	319.081	326.410	2.116	0.000	−0.054	0.000	228.999	238.774
4	**3**	N-6	202.000	0.000	−0.133	0.100	134.174	151.816	5.774	0.000	−0.079	0.147	105.329	131.126	2.344	0.000	0.158	0.000	0.000	37.604
4	**4**	N-3	260.000	0.000	0.098	0.223	47.138	85.438	6.274	0.000	0.151	0.191	35.083	68.284	2.143	0.000	−0.090	0.000	97.295	125.976
4	**5**	N-3	320.000	0.071	0.289	0.126	0.000	47.138	6.630	0.071	0.300	0.095	0.000	35.083	2.352	0.071	0.046	0.000	65.820	97.295
4	**6**	N-17	202.000	0.000	0.081	−0.146	290.643	307.411	5.774	0.000	0.030	−0.164	270.893	290.088	2.344	0.000	−0.122	0.000	125.976	147.977
4	**7**	N-17	260.000	0.077	0.042	−0.240	269.061	290.643	6.274	0.077	−0.040	−0.240	250.035	270.893	2.143	0.077	−0.226	0.000	147.977	168.086
4	**8**	N-13	156.000	0.000	−0.015	−0.090	252.167	269.061	5.236	0.000	−0.049	−0.077	225.152	250.035	2.116	0.000	0.088	0.000	332.484	360.000
4	**9**	N-10	320.000	0.071	−0.279	−0.150	191.403	225.094	6.630	0.071	−0.316	−0.002	164.921	195.942	2.352	0.071	−0.051	0.000	246.315	275.190
4	**10**	N-20	404.000	0.267	0.287	−0.214	318.376	328.265	7.130	0.267	0.210	−0.289	301.018	310.998	2.133	0.267	−0.257	0.000	219.144	228.999
4	**11**	N-10	376.000	0.133	−0.374	0.055	151.816	191.403	7.223	0.133	−0.321	0.200	131.126	164.921	2.307	0.133	0.125	0.000	275.190	303.518
4	**12**	N-25	156.000	0.091	0.085	−0.032	337.028	341.838	5.236	0.091	0.079	−0.046	326.410	333.231	2.116	0.091	−0.091	0.000	176.504	184.456
4	**13**	N-25	156.000	0.091	0.088	−0.025	341.838	346.648	5.236	0.091	0.084	−0.036	333.231	340.051	2.116	0.091	−0.090	0.000	184.456	192.408
4	**14**	N-25	156.000	0.091	0.089	−0.017	346.648	351.459	5.236	0.091	0.087	−0.026	340.051	346.872	2.116	0.091	−0.087	0.000	192.408	200.360
4	**15**	N-25	156.000	0.091	0.090	−0.010	351.459	356.269	5.236	0.091	0.090	−0.015	346.872	353.692	2.116	0.091	−0.083	0.000	200.360	208.312
4	**16**	N-13	250.000	0.154	−0.127	−0.209	225.094	252.167	6.147	0.154	−0.211	−0.124	195.942	225.152	2.227	0.154	0.182	0.000	303.518	332.484
4	**17**	N-20	448.000	0.333	0.289	−0.311	307.411	318.376	7.809	0.333	0.183	−0.383	290.088	301.018	1.632	0.333	−0.196	0.000	238.774	246.315
4	**18**	N-6	558.000	0.314	−0.163	0.453	85.438	134.174	8.292	0.314	0.027	0.480	68.284	105.329	1.759	0.314	0.298	0.000	37.604	65.820
4	**19**	N-20	202.000	0.083	0.152	−0.085	328.265	333.209	5.774	0.083	0.123	−0.123	310.998	319.081	2.344	0.083	−0.145	0.000	208.312	219.144

[a] L:level, ID: node, F: parent node, *I*: index value, *D*: arc distance, *X* and *Y*: point coordinates, α_i: initial angle, and α_f: final angle.

dx.doi.org/10.1021/ci100484z |*J. Chem. Inf. Model.* 2011, 51, 1216–1232

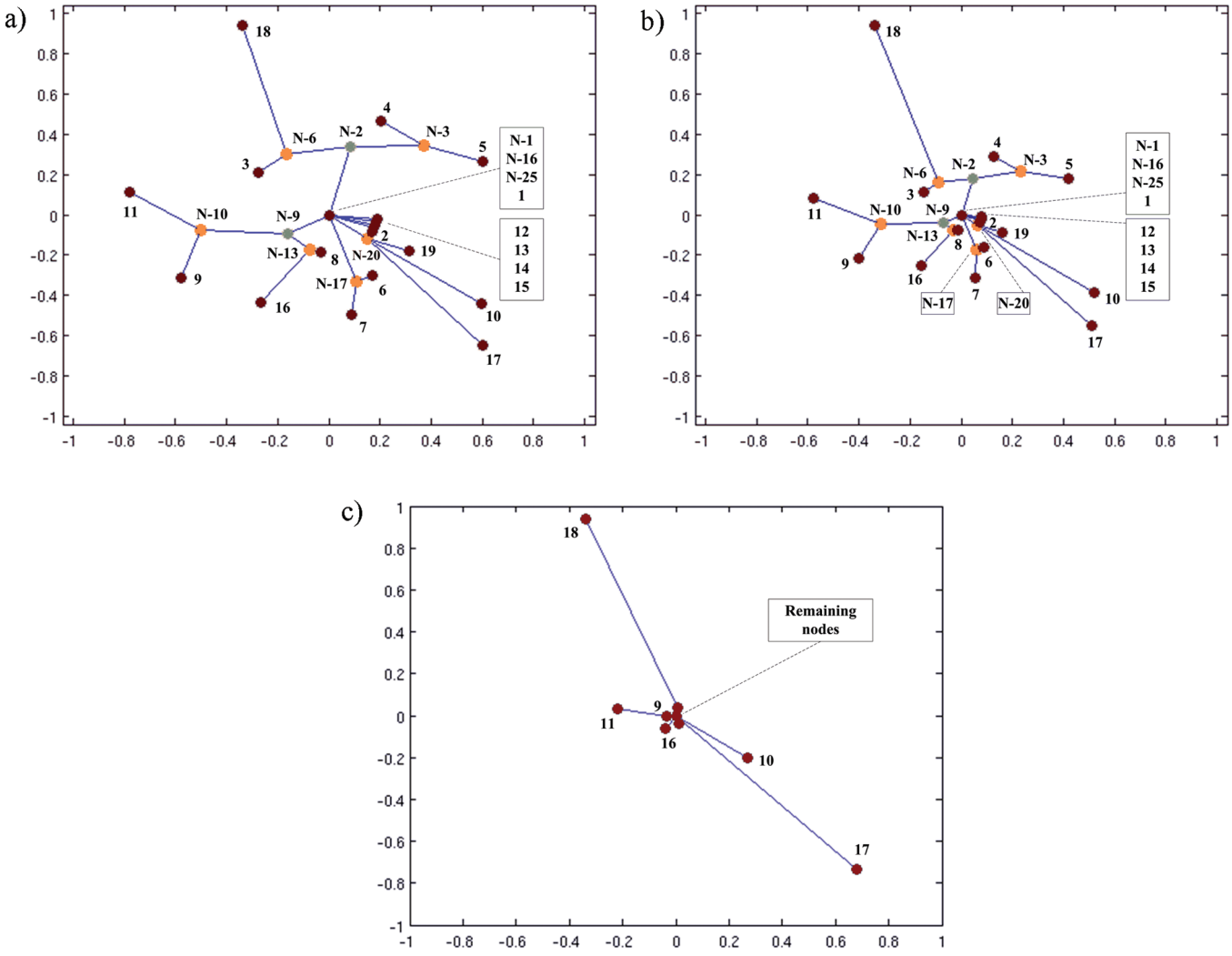


Figure 3. Behavior of the edge distance in the WMCST snowflake diagram.

to the child node with the circumference defines the points where the parent node could have been placed. The point where the child node is represented is equal to the point where this circumference intersects the bisectrix of the quarter assigned to the node. For example, for obtaining the point that displays the point N_{11}, the steps are:

(i) The bisectrix of the quarter assigned to the node N_{11} is obtained.

(ii) The intersection of this bisectrix with the circumference assigned to the node N_1 is calculated.

(iii) The circumference, with radius equal to: $d(N_1, N_{11})$ and with the center equal to the intersection point from the previous step, is obtained.

(iv) The node N_{11} could be assigned any point on the circumference arc once obtained the bisectrix (see Figure 1b).

(c) The same process is done for the other child nodes.

The process is repeated for all the levels of the tree obtained in Figure 1c, the full snowflake diagram for the WMCST example in Table 1. If the process is exhaustively analyzed, then the snowflake diagram considers the following for the assignment of the points to the different nodes of the tree: Each node is assigned to a point in the bisectrix of the circumference quarter assigned to that point, and the distance of that point to the root node is equal to the sum of the distances of the path between the root node and the given one.

The criteria chosen in the snowflake diagram for assigning the points in the bisectrix of its assigned quarter permits the use of any criteria for the calculation of distances (referred to the parent or root node as aforementioned) and the generation of a diagram that can be reproduced, to avoid problems with the representation from values of the distances lower than the minimum distance of the parent node to the bisectrix assigned to the child node and to display values of distances equal to zero keeping the characteristics of the representation.

4. EXPERIMENTAL RESULTS

Once the characteristics of the snowflake diagram have been described as well as the procedure for building it from a data set of molecules from the WMCST, how to apply this diagram will be demonstrated making use of a data set taken from the literature[42,43] and composed of 19 compounds of 2-substituted isonicotinic acid hydrazide (INH) derivates which are very effective agents in tuberculosis therapy.

Table 3. Behavior of the Distance Model on the WMCST Snowflake Diagram[a]

						$d = 1 - S$			$d = \lvert W_f - W_s \rvert$			$d = W[\mathrm{NIF}]$		
L	ID	F	W	α_i	α_f	D	X	Y	D	X	Y	D	X	Y
0	N-0	N-0	0	0.000	360.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	N-1	N-0	121	0.000	360.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	N-16	N-1	121	269.061	360.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	N-2	N-1	202	0.000	151.816	0.347	0.084	0.336	0.185	0.045	0.180	0.009	0.002	0.008
2	N-9	N-1	156	151.816	269.061	0.189	−0.163	−0.096	0.080	−0.069	−0.041	0.000	0.000	0.000
3	N-10	N-9	260	151.816	225.094	0.320	−0.503	−0.075	0.238	−0.315	−0.047	0.009	−0.009	−0.001
3	N-13	N-9	156	225.094	269.061	0.000	−0.074	−0.174	0.000	−0.031	−0.074	0.000	0.000	0.000
3	N-17	N-16	202	269.061	307.411	0.347	0.108	−0.329	0.185	0.058	−0.176	0.009	0.003	−0.008
3	N-20	N-16	156	307.411	337.028	0.189	0.149	−0.116	0.080	0.063	−0.049	0.000	0.000	0.000
3	N-25	N-16	121	337.028	360.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	N-3	N-2	260	0.000	85.438	0.160	0.372	0.344	0.133	0.234	0.216	0.009	0.013	0.012
3	N-6	N-2	202	85.438	151.816	0.000	−0.166	0.304	0.000	−0.089	0.163	0.009	−0.008	0.015
4	**1**	N-25	121	356.269	360.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	**2**	N-20	156	333.209	337.028	0.000	0.172	−0.080	0.000	0.073	−0.034	0.000	0.000	0.000
4	**3**	N-6	202	134.174	151.816	0.000	−0.277	0.209	0.000	−0.148	0.112	0.000	−0.014	0.010
4	**4**	N-3	260	47.138	85.438	0.000	0.204	0.464	0.000	0.128	0.291	0.000	0.007	0.016
4	**5**	N-3	320	0.000	47.138	0.149	0.600	0.262	0.137	0.417	0.182	0.000	0.016	0.007
4	**6**	N-17	202	290.643	307.411	0.000	0.168	−0.303	0.000	0.090	−0.162	0.009	0.008	−0.015
4	**7**	N-17	260	269.061	290.643	0.160	0.087	−0.499	0.133	0.054	−0.313	0.009	0.003	−0.017
4	**8**	N-13	156	252.167	269.061	0.000	−0.031	−0.187	0.000	−0.013	−0.079	0.000	0.000	0.000
4	**9**	N-10	320	191.403	225.094	0.149	−0.579	−0.311	0.137	−0.401	−0.216	0.009	−0.015	−0.008
4	**10**	N-20	404	318.376	328.265	0.555	0.596	−0.444	0.568	0.519	−0.387	0.078	0.062	−0.046
4	**11**	N-10	376	151.816	191.403	0.277	−0.778	0.115	0.265	−0.577	0.085	0.991	−0.989	0.146
4	**12**	N-25	156	337.028	341.838	0.189	0.177	−0.066	0.080	0.075	−0.028	0.000	0.000	0.000
4	**13**	N-25	156	341.838	346.648	0.189	0.182	−0.051	0.080	0.077	−0.022	0.000	0.000	0.000
4	**14**	N-25	156	346.648	351.459	0.189	0.186	−0.036	0.080	0.079	−0.015	0.000	0.000	0.000
4	**15**	N-25	156	351.459	356.269	0.189	0.188	−0.020	0.080	0.080	−0.009	0.000	0.000	0.000
4	**16**	N-13	250	225.094	252.167	0.320	−0.265	−0.435	0.215	−0.154	−0.252	0.810	−0.422	−0.692
4	**17**	N-20	448	307.411	318.376	0.693	0.600	−0.646	0.668	0.509	−0.548	0.233	0.158	−0.171
4	**18**	N-6	558	85.438	134.174	0.654	−0.339	0.941	0.815	−0.339	0.941	0.233	−0.085	0.235
4	**19**	N-20	202	328.265	333.209	0.173	0.316	−0.177	0.105	0.162	−0.091	0.000	0.000	0.000

[a] L:level, ID: node, F: parent node, W: Wiener index, D: arc distance, X and Y: point coordinates, α_i: initial angle, and α_f: final angle.

4.1. Analysis of the WMCST Parameters. The parameters involved in building the WMCST are the values of the weights of the nodes and the values of the distances between the nodes. From a given MCST, different WMCST could be built which would lead to different snowflake diagrams that would allow the study of the molecule data sets from different points of view.

4.1.1. Behavior of the Weight Assigned to Nodes. The weight assigned to a node in the WMCST is obtained based on the value of the molecular descriptor or property used for creating the tree. The value of a molecular descriptor is affected by the characteristics of the molecular structure, such as size, types of atoms and links, symmetry, etc.[44−45] Therefore, the type of descriptor used would define the circumference arc assigned to each node, generating diagrams where the nodes that correspond to the molecules would be more or less spread.

Figure 2 shows the MCST obtained (top) and the snowflake diagrams for the different WMCST generated (see Table 2) for the three descriptors tested: Wiener,[46] Randic,[47] and Balaban.[48] In Table 2 the corresponding experimental results are summarized. The method used for the calculation of the arc's distance has been proposed in eq 1, using Tanimoto index as the similarity index.

As can be seen in Figure 2, the MCST is structured in four levels, with the root node composed of a MCS common to all the data set that matches the structure of molecule **1**. Three substructures are the level one of the MCST (nodes N-2, N-9, and N-16), and 7 substructures compose the level 2 (nodes N-3, N-6, N-10, N-13, N-17, N-20, and N-25), where the 19 molecules of the data set are grouped in 5 groups of 2 molecules, with a set of 4 molecules and a group of 5 molecules.

As can be seen in Figure 2 and Table 2, the use of different descriptors does not impact the weight of the arcs of the WMCST generated or on the distance of the different arcs that link the nodes of the snowflake diagram. However, the descriptor impacts the arc assigned to each node and, therefore, at the point where each node would be displayed in the snowflake diagram.

For example, the node N-2 has assigned a weight of 202.0, 5.774, and 2.344 that corresponds to the descriptor of the substructure (Wiener, Randic, and Balaban). These values result in the node having a circumference arc assigned of 151.8°, 131.1°, and 125.9° respectively, because the nodes of the first level are distributed within 360° regarding the value of the descriptor. Where the Wiener index is used, the nodes N-2, N-9, and N-16

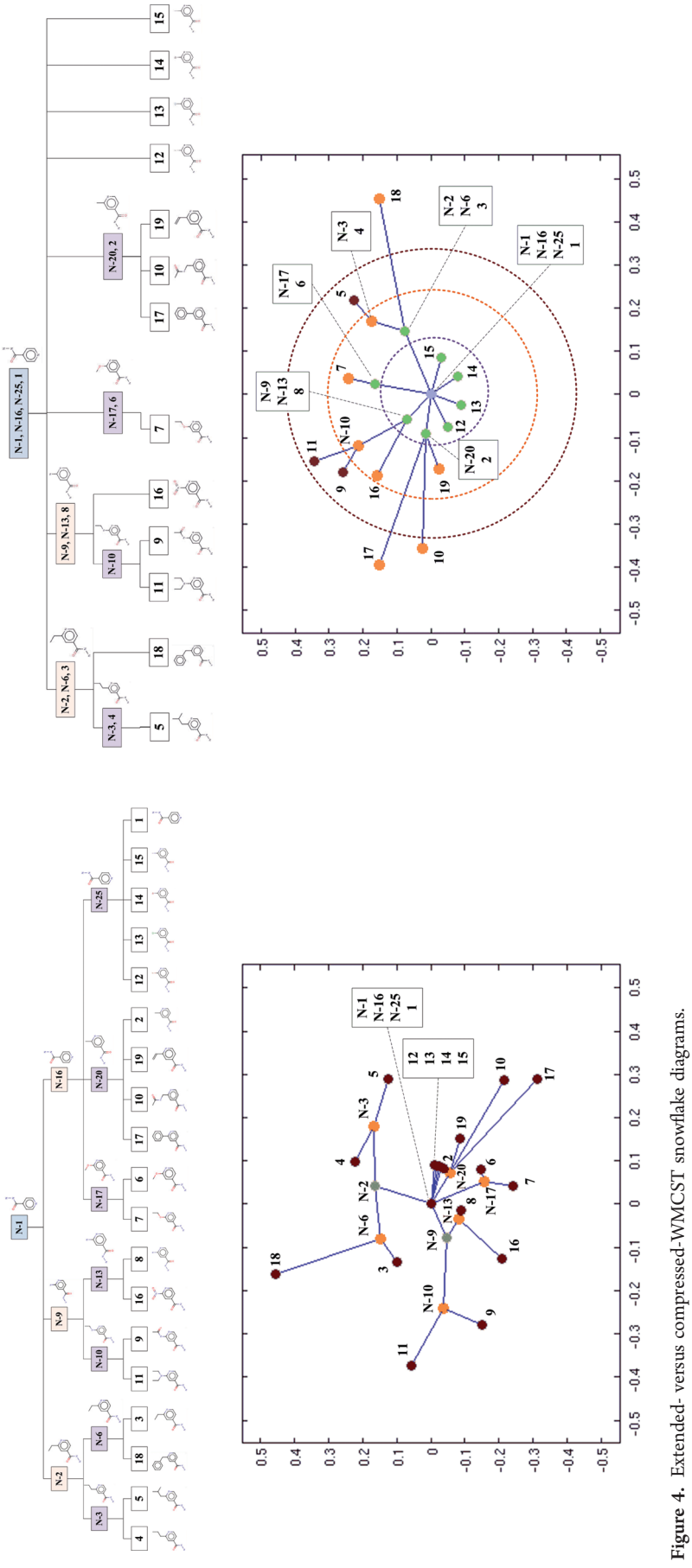


Figure 4. Extended- versus compressed-WMCST snowflake diagrams.

dx.doi.org/10.1021/ci100484z |*J. Chem. Inf. Model.* 2011, 51, 1216–1232

Table 4. Experimental Results for Data Set for Extended- and Compressed-WMCST. [a]

extended									compressed								
L	ID	F	*W*	*D*	*X*	*Y*	α_i	α_f	L	ID	F	*W*	*D*	*X*	*Y*	α_i	α_f
0	N-0	N-0	0.000	0.000	0.000	0.000	0.000	360.000	0	N-0	N-0	0.000	0.000	0.000	0.000	0.000	360.000
1	N-1	N-0	121.000	0.000	0.000	0.000	0.000	360.000	1	1	N-0	121.000	0.000	0.000	0.000	0.000	360.000
2	N-16	N-1	121.000	0.000	0.000	0.000	269.061	360.000	1	N-1	N-0	121.000	0.000	0.000	0.000	0.000	360.000
2	N-2	N-1	202.000	0.167	0.041	0.162	0.000	151.816	1	N-16	N-0	121.000	0.000	0.000	0.000	0.000	360.000
2	N-9	N-1	156.000	0.091	−0.078	−0.046	151.816	269.061	1	N-25	N-0	121.000	0.000	0.000	0.000	0.000	360.000
3	N-10	N-9	260.000	0.154	−0.242	−0.036	151.816	225.094	2	**2**	N-16	156.000	0.000	−0.090	0.014	150.448	192.358
3	N-13	N-9	156.000	0.000	−0.035	−0.084	225.094	269.061	2	**3**	N-1	202.000	0.000	0.148	0.076	0.000	54.269
3	N-17	N-16	202.000	0.167	0.052	−0.158	269.061	307.411	2	**6**	N-16	202.000	0.000	0.025	0.165	54.269	108.537
3	N-20	N-16	156.000	0.091	0.072	−0.056	307.411	337.028	2	**8**	N-1	156.000	0.000	−0.058	0.070	108.537	150.448
3	N-25	N-16	121.000	0.000	0.000	0.000	337.028	360.000	2	**12**	N-25	156.000	0.091	−0.076	−0.050	192.358	234.269
3	N-3	N-2	260.000	0.077	0.179	0.165	0.000	85.438	2	**13**	N-25	156.000	0.091	−0.023	−0.088	234.269	276.179
3	N-6	N-2	202.000	0.000	−0.080	0.146	85.438	151.816	2	**14**	N-25	156.000	0.091	0.042	−0.081	276.179	318.090
4	**1**	N-25	121.000	0.000	0.000	0.000	356.269	360.000	2	**15**	N-25	156.000	0.091	0.085	−0.033	318.090	360.000
4	**2**	N-20	156.000	0.000	0.083	−0.038	333.209	337.028	2	N-13	N-1	156.000	0.000	−0.058	0.070	108.537	150.448
4	**3**	N-6	202.000	0.000	−0.133	0.100	134.174	151.816	2	N-17	N-16	202.000	0.167	0.025	0.165	54.269	108.537
4	**4**	N-3	260.000	0.000	0.098	0.223	47.138	85.438	2	N-2	N-1	202.000	0.167	0.148	0.076	0.000	54.269
4	**5**	N-3	320.000	0.071	0.289	0.126	0.000	47.138	2	N-20	N-16	156.000	0.091	−0.090	0.014	150.448	192.358
4	**6**	N-17	202.000	0.000	0.081	−0.146	290.643	307.411	2	N-6	N-1	202.000	0.000	0.148	0.076	0.000	54.269
4	7	N-17	260.000	0.077	0.042	−0.240	269.061	290.643	2	N-9	N-1	156.000	0.091	−0.058	0.070	108.537	150.448
4	**8**	N-13	156.000	0.000	−0.015	−0.090	252.167	269.061	3	**4**	N-2	260.000	0.000	0.170	0.174	37.020	54.269
4	**9**	N-10	320.000	0.071	−0.279	−0.150	191.403	225.094	3	**7**	N-17	260.000	0.077	0.036	0.241	54.269	108.537
4	**10**	N-20	404.000	0.267	0.287	−0.214	318.376	328.265	3	**10**	N-20	404.000	0.267	−0.357	0.023	168.262	184.326
4	**11**	N-10	376.000	0.133	−0.374	0.055	151.816	191.403	3	**16**	N-13	250.000	0.154	−0.188	0.157	129.903	150.448
4	**12**	N-25	156.000	0.091	0.085	−0.032	337.028	341.838	3	**17**	N-20	448.000	0.333	−0.397	0.150	150.448	168.262
4	**13**	N-25	156.000	0.091	0.088	−0.025	341.838	346.648	3	**18**	N-6	558.000	0.314	0.456	0.153	0.000	37.020
4	**14**	N-25	156.000	0.091	0.089	−0.017	346.648	351.459	3	**19**	N-20	202.000	0.083	−0.172	−0.025	184.326	192.358
4	**15**	N-25	156.000	0.091	0.090	−0.010	351.459	356.269	3	N-10	N-9	260.000	0.154	−0.120	0.214	108.537	129.903
4	**16**	N-13	250.000	0.154	−0.127	−0.209	225.094	252.167	3	N-3	N-2	260.000	0.077	0.170	0.174	37.020	54.269
4	**17**	N-20	448.000	0.333	0.289	−0.311	307.411	318.376	4	**5**	N-3	320.000	0.071	0.220	0.225	37.020	54.269
4	**18**	N-6	558.000	0.314	−0.163	0.453	85.438	134.174	4	**9**	N-10	320.000	0.071	−0.181	0.259	120.080	129.903
4	**19**	N-20	202.000	0.083	0.152	−0.085	328.265	333.209	4	**11**	N-10	376.000	0.133	−0.156	0.345	108.537	120.080

[a] L:level, ID: node, F: parent node, *W*: Wiener index, *D*: arc distance, *X* and *Y*: point coordinate, α_i: initial angle, and α_f: final angle.

have values of 202.0, 156.0, and 121.0, respectively, so they are proportionally distributed within the 360° of the circumference assigned to the parent node (N-1) in the following way: 151.8, 117.2, and 91.0 respectively. Besides, in the case of the Balaban index, the nodes N-2, N-9, and N-16 show weights of 2.344, 2.116, and 2.240, which means circumference arcs assigned of 126.0, 113.7, and 120.3 respectively. As can be observed, the use of different descriptors means that the changes in the percentage of the circumference arc assigned to the nodes of the same level are substantial: 42.17, 32.6, and 25.2% for Wiener and 35.0, 31.6, and 33.4% for Balaban.

The use of different descriptors results in a different distribution of the nodes in the circular space of the snowflake diagram. Even when the distance of each node to the parent node and to the center of the coordinates does not change, the Euclidean distance between the points where the nodes are displayed does.

As can also be observed in Figure 2, the use of different descriptors impacts the sorting of the nodes in the snowflake diagram. In the representation for the Wiener and Randic descriptors, the nodes are distributed in the same clockwise order. However, with Balaban index, the sorting of the nodes is different. This is due to the value of the indexes (weight) of each node (substructure), as the representation is performed in ascending order for the weights of the nodes in each level.

For example, in level 1, the nodes N-2, N-9, and N-16 have a weight for the Wiener and Randic indexes in the order N-2 > N-9 > N-16, while for the Balaban index, the weights are N-2 > N-16 > N-9. If we now focus on the child nodes of the N-2 node, nodes N-3 and N-6, then it can be seen that the weights for the Wiener and Randic indexes are sorted like this N-3 > N-6, while for the Balaban index it is N-6 > N-3. And, finally, if we focus on the child nodes of node N-6, the molecules **3** and **18**, it can be seen that the order of the weights for Wiener and Randic index is **18** > **3**, while for Balaban it is **3** > **18**.

Therefore, the weight of the nodes of the WMCST defines their spatial distribution within the snowflake diagram, defining both the circumference arc or the circular space assigned for its representation, and the assignment order of this space.

The diagram in Figure 2 shows clearly the distribution of the different substructures of the WMCST and the molecules of the data set and, therefore, the structural characteristics of the data set. For example, it can be observed that nodes N-1, N-16, and

N-25 and molecule **1** have the same structure (and all of them are displayed in the center of coordinates), which results in the fact that the group of molecules **12**−**15** are displayed quite close among them and to the center of the circumference. Another characteristic of the diagram can be seen for molecule **3**, with the same structure of nodes N-2 and N-6, which group other sets of molecules. As can be observed, nodes N-2 and N-6 and molecule **3** are represented at the same distance with respect to the center of the circumference, while the rest of molecules (and WMCST nodes), depending on node N-2 (N-3, 4 y 5), are displayed at different distances, allowing clear inference of the structural differences of the molecules regarding the method used for the calculation of the weights of the WMCST.

4.1.2. Behavior of the Distance Assigned to Edges. In a WMCST the distance between the nodes (parent and child) is a measure of the dissimilarity between the molecular substructures represented by those nodes. As described in the previous section, the distance could be calculated following different criteria with the aim of obtaining a measure of the cost of reaching a child node from a parent node of the tree.

Figure 3 shows the snowflake diagrams obtained for the data set, considering three different criteria for the weights calculation (distance) of the WMCST arc: (a) 1 minus the similarity between the parent and child nodes (eq 1), (b) the difference of the value of the Wiener index (any other descriptor could be used) between the parent and child nodes, and (c) the Wiener index value for the nonisomorphic value existing between the molecular structures represented by the parent and child nodes (eq 2).

For a better analysis of the results, the distances have been normalized in all cases (see Table 3), making the greater distance between any child node (a molecule of the data set) to the root node of the WMCST equal to 1.

Logically, the arc weighting method does not impact the distribution of the nodes in the diagram, the weight of the nodes, or the arc assigned to each node. The method only impacts the representation distance of the nodes with regards to the center of coordinates or root node. Figure 3 and Table 2 show the results obtained, where it can be seen that the method based on similarity (Figure 3a) is more sensitive than the one based on the difference of the value of the descriptor (Figure 3b), obtaining greater values for the distance of the nodes to the center (made by the MCS structure common to all the data set that matches molecule **1**) and therefore a more accurate characterization of the differences among the elements of the data set.

Thus, the distance between the child and parent nodes is always higher in case of using eq 1 when calculating the weights of the arcs except when: (a) the distance is zero, that is, the structure of the child node is the same structure as the parent node. For example, nodes N-6 and N-2, nodes **3** and N-6, etc; and (b) in the case of molecule **18** (in a lesser extent for molecule **10**) that has values $d = 1 - S = 0.654$ and $d = |W_f - W_s| = 0.815$, due to the fact that the distance to its parent (node N-6) and, therefore, from N-2 to its parent (node N-1) is remarkably greater in the first case (0.347 vs 0.185).

This effect could provide useful information about the fact that for both molecules of the data set, the relationship between the Wiener index and the similarity shows great differences with regards to the other elements of the data set. This effect was previously described in literature.[42]

The structural differences of the molecules from the data set can be clearly seen when the value of the descriptor for the nonisomorphic fragment (Figure 3c) is used as distance. Most of the molecules have quite similar structures, those being grouped in the center of the graph because the value of the distances is close to zero. However, it could be seen that molecules **10** and **11** and specifically **17** and **18** show a high structural dissimilarity compared to the other molecules of the data set when the Wiener index is used. So, the use of the eq 2 shows a greater sensitiveness to determine dissimilarities in the data set originated, in this case, by the structural differences the Wiener index is based on.

4.1.3. Extended versus Compressed WMCST. Because of the MCST is a hierarchical and balanced structure built by molecular structures common to the molecules of a given data set, some redundancies can appear within the nodes of the MCST of different levels. If we focus on the MCST in Figure 4 (top-left), it can be seen that the structure of molecule **1** appears in different levels (nodes N-1, N-16, and N-25), a method that is used by the algorithm that builds the MCST for creating a fully balanced tree, from which the WMCST is generated. This tree could be named extended-WMCST. This characteristic results in obtaining, during the process of building the extended-WMCST, distance values equal to zero between nodes of different levels in the same branch, i.e., among the nodes N-1, N-16, N-25 and **1** or among the nodes N-9, N-13, and 8, among others, as can be seen in Figure 3.

The display of the extended-WMCST (Figure 4 top-left) shows the existence of those equal structures, as all the nodes corresponding to those structures are displayed at the same distance from the center of the circumference.

However, for the analysis of the characteristics and diversity of the data set, it would be better to discard the redundancies in the MCST during the process of building the WMCST and its graphic representation. Figure 4 (top-right) shows the structure of the compressed-WMCST once the redundancies are discarded.

As can be seen, nodes N-1, N-16, and N-25 are unified (disposed) together with molecule **1** which moves to the first level in the tree in the extended-WMCST. Nodes N-2 and N-6 are unified with molecule **3**, which moves from level 4 to 2, resulting in moving molecule **18** from level 4 to 3, etc.

The effect of removing redundant structures for building the compressed-WMCST can be seen in Table 4 for the data set of the 2-substituted INH derivatives when considering the Wiener index in the nodes weighting and eq 1 for the distances or arcs' weight calculation. Figure 4 shows the extended- and compressed-WMCST for the experimental and corresponding snowflake diagrams.

As can be seen in Figure 4, results are noticeably different for the extended- and compressed-WMCST. The removal of redundancies creates an unbalanced tree where molecules of the data set are in middle nodes within the tree and produce a new sorting of the nodes based on their weight. Besides, in the compressed-WMCST there are no distances equal to zero among the nodes; there not being any nodes with the same radius, which simplifies the diagram.

The simplification of the diagram for the compressed-WMCST shows other characteristics of the data set, as, for example, the similarity of the molecules **1** and **12**−**15** or molecules **6** and **7**. Hence, it can be clearly seen, i.e.:

- Diversity between molecules **6** and **7** is based on including a substituent (−CH_2−) which creates a structural distance difference of just 0.077 (see Table 4).
- Diversity among molecules **1** and **12**−**15** is also due to a substituent that creates structural differences in the distance value of 0.091 for all the molecules. This means that

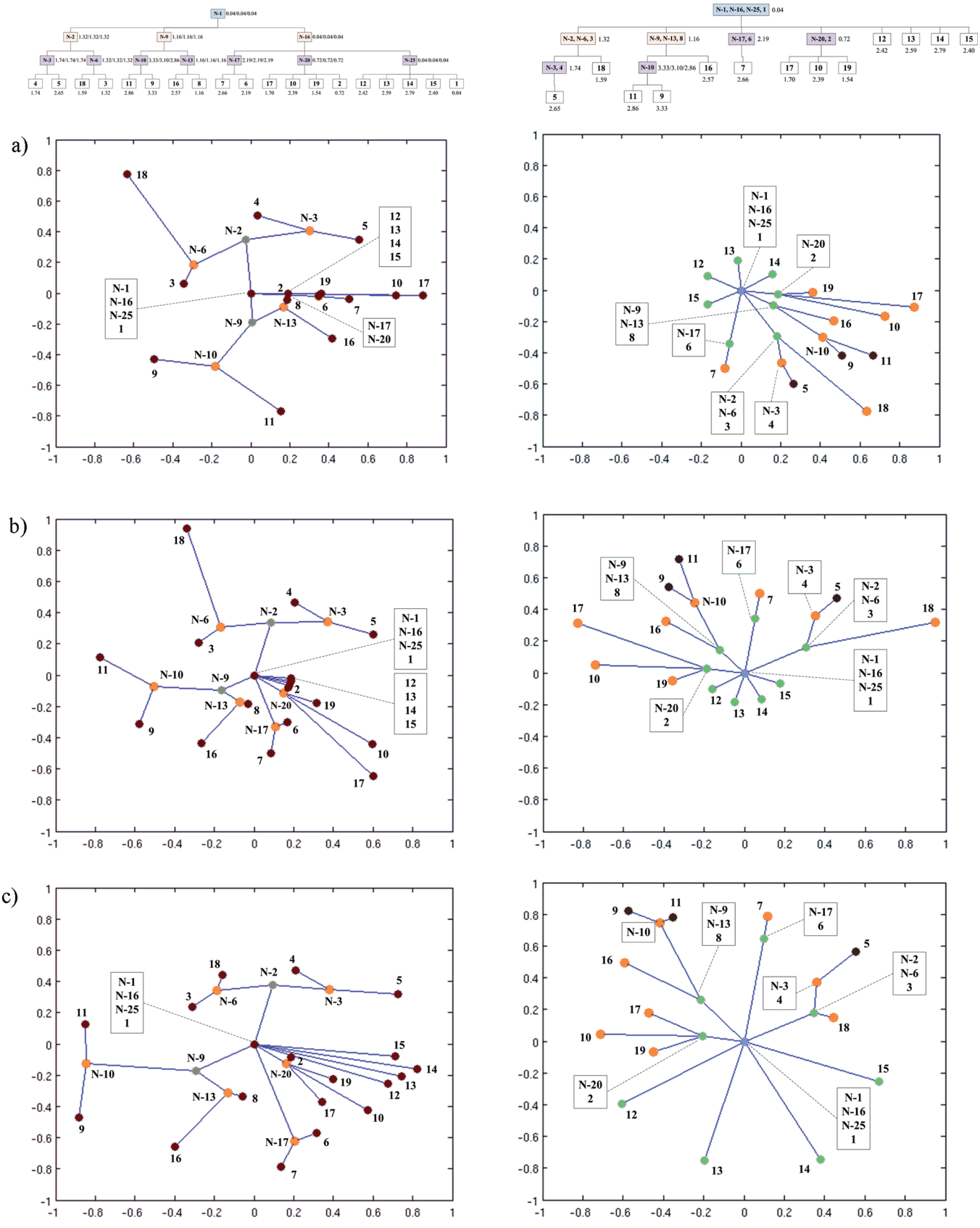


Figure 5. Behavior of −log(1/MIC) property values for the nodes weighted and distance calculation against Wiener index based node weighted and similarity based distance calculation.

dx.doi.org/10.1021/ci100484z |*J. Chem. Inf. Model.* 2011, 51, 1216–1232

the substituents are different with equal size and, as can be observed in Table 3, with equal weight (Wiener index), which creates a proportional distribution in the influence arc assigned to molecule **1.**

- Molecule **18** contains molecule **3** as a substructure but includes a substituent which results in a large structural difference (distance equal to 0.314).

In the compressed-WMCST in Figure 4, both minimal and noticeable differences among the molecules groups in the same WMCST can be seen, so structural dissimilarities can be clearly seen. The arc distribution for molecules **12**−**15** is equal and quite similar to the one for molecules **2** and **8**, displaying its structural similarity. This fact is tested as they are displayed in quite similar distances to the origin.

So, it is possible to easily spot molecule groups or clusters if we represent concentric areas that refer to the center of coordinates. Those area group molecules that are quite similar (eq 1 has been used for weighting the arcs). Its distribution within the circular area allows us to perceive the behavior of the descriptor used for weighting the nodes under analysis.

4.1.4. Molecule Property as Weighted Factor of the Nodes of WMCST. The model for building the WMCST described in this paper allows the researchers to consider any characteristic of the molecular structures represented by the nodes in the MCST as a weighting factor of themselves. As aforementioned, the molecular descriptor could be conveniently used for this purpose, as the corresponding snowflake diagram provides a view over the descriptor (or descriptors) behavior used over the molecules in the data set.

This flexibility allows us to use any characteristic for weighting the nodes of the MCST and can be used for studying the characteristics of the data set with regards to the property under analysis, with the aim of, for example, analyzing (a priori) the nature of the data set for the development and the creation of QSPAR/QSAR predictive models for that property.

In Figure 5, the snowflake diagrams corresponding to the WMCST generated for the data set of the 2-substituted INH derivatives aforementioned are shown, considering:

- The weight of the arcs, or distance between nodes, has been obtained using eq 1 and the Tanimoto similarity index. Values have been normalized for a better comparison of the results.
- The weight of the nodes from the WMCST corresponds to the value of −log(1/MIC) (minimum inhibitory concentration) of the molecules from the data set with regards to their antituberculosis activity.[42,43]
- For those nodes whose structures do not correspond to molecules of the data set and the property value is unknown, this value could be close to the maximum, average, or minimum value of the property for the child nodes.

For the data set used, there is no difference when the weight of the nodes corresponding to the structures outside the data set is close to the maximum, average, or minimum value of the child nodes. The reason for this is the fact that the node N-10 is the only one that does not correspond to any molecule of the data set. However its maximum, average, and minimum weight is higher than any of the weights of the other nodes from the same level; therefore the distribution of the nodes in the diagram does not change even when the arc assigned to node N-10 changes slightly. In Figure 5 the average value of the property values for those nodes which structure does not correspond to any of the molecules of the data set that has been used.

If we compare the diagrams corresponding to considering the property (Figure 5a) and the Wiener molecular descriptor (Figure 5b) as weighting factor for the nodes, some valuable information can be extracted. As can be observed, no extreme difference was seen in the distribution of nodes N-2 and N-9 and their child nodes (molecules **4**, **5**, **18**, **3**, **11**, **9**, **16**, and **8**), but a difference does exist in the nodes and the molecules depending on node N-25. The reason for this is that the structure of node N-25 is the same as the one for node N-16 and molecule **1** of the data set and that molecule is the MCS common to the whole data set having also the smallest value of the property of the set −log(1/MIC) = 0.04. This explains why node N-16 is assigned a very small arc (5.71°), as it has a very small percentage of the value of the property (1.587%) among the first level nodes.

So, the child nodes must be distributed in a very small arch, practically being all over the same line. It can be observed that the nodes corresponding to molecules **12**−**15** are almost at the same point, which proves that all those nodes have similar property values. This behavior can also be seen for other nodes: N-17, N-20, 2, etc.

If we observe the compressed-WMCST (Figure 5a and b), then we can clearly noticed the distribution of nodes **12**−**15**, which are at the same distance from the center (equal similarity with molecule **1**) and have assigned a similar circumference arc, which happens because their similarity values are quite close.

When the Wiener index is considered as the nodes' weighting factor, the effect is the opposite, that is, the molecules more similar to molecule **1** have lower weight values than those molecules which are more different; this is due to a greater difference in size (topological aspect considered in the Wiener index). Thus, molecule **17** with a Wiener value equal to 448 and −log(1/MIC) = 1.70, has a circumference arc assigned of 17.81°, and molecule **10** with a Wiener value equal to 16.07 and −log(1/MIC) = 2.39 has an arc of 16.07°. It is observed that when incrementing the value of the property under analysis, the similarity with molecule **1** also increases, decreasing the distance with regards to the center of the diagram and the Wiener index value and therefore the Euclidean distance to other nodes.

This proper behavior of the Wiener index for the study of the property in this data set has already been observed by Bagchi et al.[42] who proposed MIC predictive models based on this descriptor.

Observing Figure 5b and c the relationship between the value of the property and the similarity can be studied. As can be appreciated again, molecules **12**−**15** have higher similarity values with regards to molecule **1**, being at a lesser distance from the center (Figure 5b), however they have property values quite higher than the value for molecule **1** and some of the highest of the whole data set, being displayed quite far from the center in Figure 5c. It can be observed that generally molecules with the higher property value are more alike to molecule **1** and vice versa.

For example, molecule **17** is placed at a higher distance from the center (is less similar to molecule **1**) than molecule **10**, even when the property value for molecule **17** is smaller [-log(1/MIC) = 1.70] than the value for molecule **10** [-log(1/MIC) = 2.39], which happens also with the value of the arc assigned to both molecules (5.02° for **17** and 7.06° for **10**).

Figure 5 allows us to clearly observe the deviations that certain molecules of the data set have with regards to the general behavior of the data set aforementioned. For example, molecule **2** shows a relative low property value (0.72) when considering its similarity with molecules **12**−**15**, which tells us that the halogen

substituents are responsible for increasing the property value. Furthermore, molecule **18**, quite different from molecule **1**, has a relative low property value (similar to the one of molecules **17** and **19**), totally different from other molecules of the data set (for example **9** or **11**). This behavior helps us to appreciate the limited effect of benzene rings on the value.

Thus, previous studies[42,43] detect different groups of molecules in data set studied with different characteristics and behavior. Authors for the development of QSAR models eliminate the group composed of molecules **8**−**11** (compounds having amino functions) and the group composed of molecules **10** and **17**−**19** (compounds having phenyl, vinyl, and methylene groups) due to their different behavior with the remaining molecules of the data set.

This data set behavior has been clearly shown (graphically and numerically) in the previous WMCST snowflake diagram, and it can be found if we represent in a same diagram the compressed WMCST of Figure 5b and c.

In Figure 6 we have considered the Wiener index as node weight and: (a) similarity value following eq 1 for the calculation of edge distances (blue nodes) and (b) −log(1/IMC) increment for the calculation of edge distances (red, gray, and yellow nodes). Thus, we observe groups of molecules with different behavior:

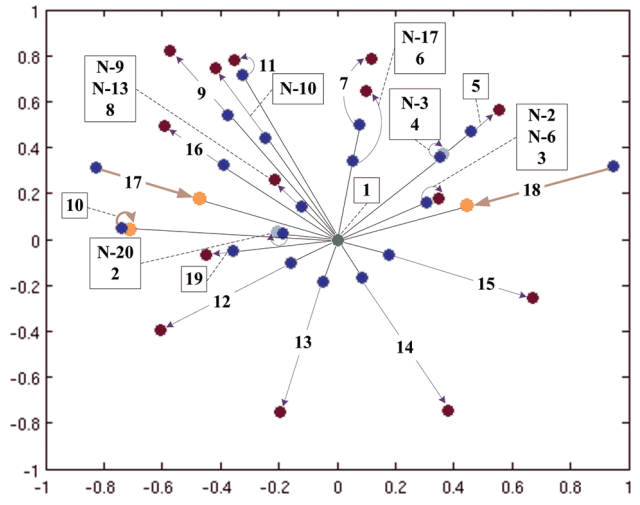


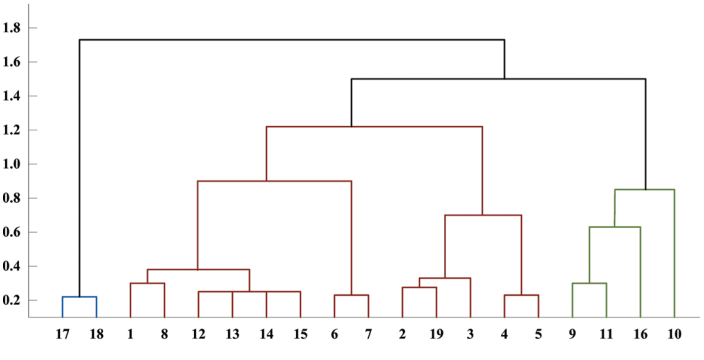
Figure 6. Combining WMCST snowflake diagram to visualize the data set behavior againts −log(1/MIC) values, Wiener index, and similarity.

- For a set of molecules there exists a high and positive distance between blue and red nodes. This is the set composed by molecules **12**−**15**. This set corresponds with the halides derivates of molecule **1**. These molecules are very similar to molecule **1**, presenting, however, a slightly higher value of the property. Moreover, for molecule **2**, also very similar to molecule **1**, this increment is quite small (the node is represented in gray). All these molecules share the same circular cluster in Figure 4b, however, we can be observe the different behavior of the halides derivates versus the short aliphatic chain derivates, that is also observed for molecules **3**−**5**.
- Another set of molecules (red colored) shows a similar positive increment: **6**−**9**, **11**, and **16**. They are molecules with short R-derivatives including N and O atoms and double links.
- Finally, we observe the anomalous behavior of molecules **10**, **17**, and **18**. These molecules have a negative increment (theses nodes have been painted in yellow). This negative increment is very low for molecule **10** with a substituent close to the previous group, while molecules **17** and **18** present cycles, confirming that cyclic substituents (the more dissimilar molecules from molecule **1**) result in low property values.

In this figure we can easily observe the relationship between the structural similarity and the property as well as the influence of the characteristics of 2-substituent to the property behavior. Similar analysis could be performed using different descriptors, therefore obtaining information of the relationship between the molecular descriptors and the property.

Figure 7 shows the results for principal component analysis (PCA)[5] and hierarchical clustering analysis (Ward clustering)[8] for the data set studied using structural similarity (MCS based) as representation space. As we can observe, both Ward and PCA clustering methods corroborate the results obtained although providing less fine information about the data set characteristics than the snowflake diagrams.

Although Ward (Figure 7 left) and MCST (Figure 4) methods generate a hierarchical classification of the data set and some equal clusters are obtained (i.e., molecules **12**−**15**, **4** and **5**, **6** and **7**, etc.), the complete results are quite different. On the one hand the MCST method groups molecules over the MCS generating new nodes corresponding to the common MCS found. On the



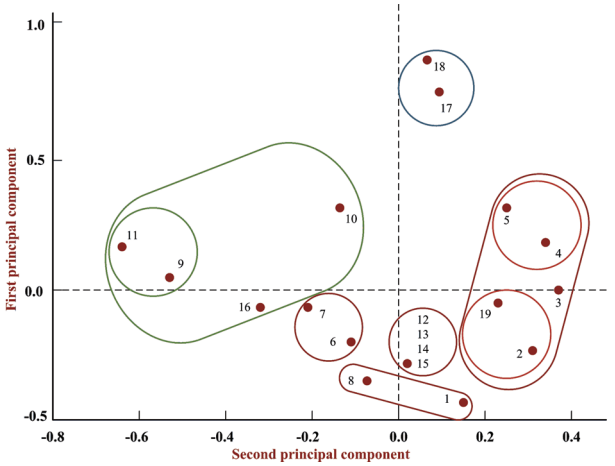


Figure 7. Comparison with other clustering methods: Ward (left) clustering and PCA (right) analysis of the data set.

dx.doi.org/10.1021/ci100484z |J. Chem. Inf. Model. 2011, 51, 1216–1232

other hand, the Ward method uses the distance, based on the similarity matrix of the data set, for grouping molecules.

The Ward method finds the different behavior of molecules **17** and **18** generating a specific cluster for both molecules. Besides, molecule **10** is classified as a singleton showing the characteristical behavior of this molecule. As in the snowflake diagram halides derivates (molecules **12**−**15**) are classified in the same cluster which is grouped with molecules **1** and **8** in a higher cluster. If we see the snowflake diagram of Figure 6, these molecules are included in the inner circle closer to the center of the diagram. However, the different behavior of molecules **12**−**15** against **8** is not appreciated in Ward clustering, as shown in Figure 6. Some clusters as the composed of molecules **4** and **5** or **6** and **7** are also appreciated in Figure 6; molecules are in the same arc of the diagram, showing a positive increment, although in Figure 6 is possible to observe the different behavior of molecule **4** against molecule **5**.

When first against second principal components of PCA analysis is represented (Figure 7-right) we cannot observe a clear classification of the data set. When we compare the clusters obtained with the Ward method against PCA analysis, we observe disperse clusters that are hardy recognizable. Although the specific behavior of groups of molecules such as: **10** or **17** and **18**, or **12**, **13**, **14** and **15** is shown, other characteristics of the data set discovered by the snowflake diagrams are not observed.

5. DISCUSSION

Molecular data sets fragmentation for its later analysis and QSAR models building is a method that has revealed excellent results in recent years. This technique allows us to associate information about the fragments that constitute the compounds to the chemical databases as well as to build structures for a more efficient molecules selection in the development of predictive models.

The first step needed in this process is to determine which molecules and variables are the most appropriate for not detecting degeneracies (due to a rare behavior of the molecule) during the building process of the model. This is a long and tedious process that uses different computational and graphical solutions, for which, the snowflake diagrams proposed in this paper fit perfectly.

Snowflake diagrams based on the WMCST representation permit us to easily notice the data set behaviors as well as the deviations in the molecules of the data set regarding the parameters under analysis. These diagrams permit the use of any descriptor (or a combination of more than one), models, similarity indexes, and molecules' properties for generating different representations that offer graphical and numerical information about the behavior of the data set to the researcher.

The advantages of snowflake diagrams versus other classical methods are based on the flexibility the researcher has in the survey and study of the data set characteristics. Thus, WMCST, and therefore the corresponding snowflake diagram, can be built using different distance and weighting models. Different similarity measurements (any model and metric) as well as molecular descriptors can be considered, and visual and numerical analysis can be performed comparing the behavior of the data set regarding the property or activity studied. This analysis allows the researcher to find any anomalous behavior of molecules or groups of molecules; these behaviors latter detected as cliffs in the building of prediction models.

Furthermore, the process for building MCST, WMCST, and the snowflake diagram is quick. It allows the researcher to extract extra information like nonisomorphic substructures and to analyze the behavior of these nonisomorphic fragments versus the property values of similar molecules of the data set.

In this paper, we have presented the basis of the snowflake diagrams based on WMCST, demonstrating its use in the study of a 2-substituted isonicotinic acid hydrazide derivates data set behavior regarding different descriptor and similarity measures. The results clearly show the behavior of the molecules and the deviation that some elements in the data set have when related to the Wiener index with regards to the property −log(1/MIC), as previously observed in other papers. The information that can be obtained permits the researcher to determine the behavior of other groups of molecules, allowing the researcher to select clusters and to observe deviations that might appear in the molecules of the data set in QSAR applications.

In future papers we will show how the information gathered with these diagrams can be used in different aspects of computational chemistry. On the one hand, the distribution of the nodes around the center of coordinates, representing the MCS, allows us to choose the clusters on the basis of distances or concentric circles. On the other hand, the position of the different fragments or molecules in the diagram allows the researcher to assign them predefined variables based on its spatial position and the later calculation of distance that might only be used together with the other parameters (descriptors, similarity, etc.) for the development of predictive models for the physic−chemical properties as well as biological activity.

AUTHOR INFORMATION

Corresponding Author

*E-mail: iluque@uco.es. Telephone: +34-957-212082.

REFERENCES

(1) Balaban, A. T. *Chemical Applications of Graph Theory*; Academic Press: New York, 1976.

(2) Gross, J.; Yellen, J. *Graph Theory and Its Applications*, 2nd ed.; Chapman and Hall/CRC Press: Boca Raton, FL, 2005.

(3) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Springer: Dordrecht, The Netherlands, 2003.

(4) Cerruela García, G.; Luque Ruiz, I.; Gómez-Nieto, M. A. Representation of the Molecular Topology of Cyclical Structures by Means of Cycle Graphs. 1. Extraction of Topological Properties. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 447–461.

(5) Jain, A. K.; Murty, M. N.; Flynn, P. J. Data Clustering: A Review. *ACM Comput. Surveys* **1999**, *31* (3), 264–323.

(6) Liu, W.; Johnson, D. E. Clustering and Its Application in Multitarget Prediction. *Curr. Opin. Drug Discovery Dev.* **2009**, *12* (1), 98–107.

(7) Maggiora, G. F. On Outliers and Activity Cliffs - Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535–1535.

(8) Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244.

(9) Holliday, J. D.; Rodgers, S. L.; Willett, P.; Chen, M. Y.; Mahfouf, M. Clustering Files of Chemical Structures Using the Fuzzy K-Means Clustering Method. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 894–902.

(10) Luque Ruiz, I.; Cerruela Garcia, G.; Gomez-Nieto, M. A. Clustering Chemical Databases Using Adaptable Projection Cells and MCS Similarity Values. *J. Chem. Inf. Model.* **2005**, *45* (5), 1178–1194.

(11) Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O. A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models. *J. Chem. Inf. Model.* **2005**, *45*, 839–849.

(12) Hawkins, D. M.; Basak, S. C.; Shi, X. QSAR with Few Compounds and Many Features. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 663–670.
(13) Zheng, W.; Tropsha, A. Novel Variable Selection Quantitative Structure-Property Relationship Approach Based on the k-Nearest-Neighbor Principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185–194.
(14) Liu, R.; Zhou, D. Using Molecular Fingerprint as Descriptors in the QSPR Study of Lipophilicity. *J. Chem. Inf. Model.* **2008**, *48*, 542–549.
(15) Clark, A. M. 2D Depiction of Fragment Hierarchies. *J. Chem. Inf. Model.* **2010**, *50*, 37–46.
(16) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold TreesVisualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **2007**, *47*, 47–58.
(17) Wawer, M.; Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Structure-Activity Relationship Anatomy by Network-like Similarity Graphs and Local Structure-Activity Relationship Indices. *J. Med. Chem.* **2008**, *51*, 6075–6084.
(18) Gardiner, E. J.; Gillet, V. J.; Willett, P.; Cosgrove, D. A. Representing Clusters Using a Maximum Common Edge Substructure Algorithm Applied to Reduced Graphs and Molecular Graphs. *J. Chem. Inf. Model.* **2007**, *47*, 354–366.
(19) Lounkine, E.; Wawer, M.; Wassermann, A. M.; Bajorath, J. SARANEA: A Freely Available Program to Mine Structure-Activity and Structure-Selectivity Relationship Information in Compound Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 68–78.
(20) Lepp, Z.; Huang, C.; Okada, T. Finding Key Members in Compound Libraries by Analyzing Networks of Molecules Assembled by Structural Similarity. *J. Chem. Inf. Model.* **2009**, *49*, 2429–2443.
(21) Tanaka, N.; Ohno, K.; Niimi, T.; Moritomo, A.; Mori, K.; Orita, M. Small-World Phenomena in Chemical Library Networks: Application to Fragment-Based Drug Discovery. *J. Chem. Inf. Model.* **2009**, *49*, 2677–2686.
(22) Randic, M.; Zupan, J.; Vikic-Topic, D. On Representation of Proteins by Star-like Graphs. *J. Mol. Graphics Modell.* **2007**, *26*, 290–305.
(23) Boitmanis, K.; Brandes, U.; Pich, C. Visualizing Internet Evolution on the Autonomous Systems Level. *Lect. Notes Comput. Sci.* **2007**, *4875*, 365–376.
(24) Stasko, J.; Catrambone, R.; Guzdial, M.; McDonald, K. An Evaluation of Space Filling Information Visualizations for Depicting Hierarchical Structures. *Int. J. Hum.-Comput. Stud.* **2000**, *53* (5), 663–694.
(25) Agrafiotis, D. K.; Bandyopadhyay, D.; Farnum, M. Radial Clustergrams: Visualizing the Aggregate Properties of Hierarchical Clusters. *J. Chem. Inf. Model.* **2007**, *47*, 69–75.
(26) Brady, N.; Riley, T.; Short, H. *The Geometry of the Word Problem for Finitely Generated Groups*; Birkhäuser Verlag: Basel, Switzerland, 2007.
(27) Ponniah, P. *Data Warehousing Fundamentals for IT Professionals*; John Wiley & Sons: Hoboken, New Jersey, 2010.
(28) Raymond, J. W.; Willett, P. Maximum Common Subgraph Isomorphism Algorithms for the Matching of Chemical Structures. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 521–533.
(29) Cerruela García, G.; Luque Ruiz, I.; Gómez-Nieto, M. A. Step-by-Step Calculation of All Maximum Common Substructures through a Constraint Satisfaction based Algorithm. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (1), 30–41.
(30) Chaudhaery, S.; Roy, K.; Saxena, A. K. Consensus Superiority of the Pharmacophore-Based Alignment, Over Maximum Common Substructure (MCS): 3D-QSAR Studies on Carbamates as Acetylcholinesterase Inhibitors. *J. Chem. Inf. Model.* **2009**, *49* (6), 1590–1601.
(31) Varmuza, K.; Penchev, P. M. Maximum Common Substructures of Organic Compounds Exhibiting Similar Infrared Spectra. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 420–427.
(32) Vargyas, M.; Csizmadia, F. Hierarchical clustering of chemical structures by maximum common substructures. In Proceedings of the ICCS International Conference of Chemical Structures, Noordwijkerhout, The Netherlands, June 1−5, 2008; ICCS: Noordwijkerhout, The Netherlands, 2008.
(33) *JChem*, version 5.3.7; Chemaxon Ltd.: Budapest, Hungary; http://www.chemaxon.com. Accessed October 10, 2010.
(34) *JRE, Java runtime Environment*, version 6.23; Oracle: Redwood Shores, CA ; http://www.java.com/. Accessed October 10, 2010.
(35) Luque Ruiz, I.; Gómez-Nieto, M. A. A Tool for the Calculation of Molecular Descriptors in the Development of QSAR Models. *Lect. Notes Comput. Sci.* **2008**, *5072*, 986–996.
(36) Luque Ruiz, I.; Gómez-Nieto, M. A. A Java Tool for the Management of Chemical Databases and Similarity Analysis Based on Molecular Graphs Isomorphism. *Lect. Notes Comput. Sci.* **2008**, *5102*, 369–378.
(37) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*, 2nd ed.; Wiley-VCH: Weinheim, Germany, 2009.
(38) Urbano Cuadrado, M.; Luque Ruiz, I.; Gómez-Nieto, M. A. A Steroids QSAR Approach Based on Approximate Similarity Measurements. *J. Chem. Inf. Model.* **2006**, *46*, 1678–1686.
(39) Urbano Cuadrado, M.; Luque Ruiz, I.; Gómez-Nieto, M. A. QSAR Models Based on Isomorphic and Nonisomorphic Data Fusion for Predicting the Blood Brain Barrier Permeability. *J. Comput. Chem.* **2007**, *28*, 1252–1260.
(40) Luque Ruiz, I.; Urbano Cuadrado, M.; Gómez-Nieto, M. A. Data Fusion of Similarity and Dissimilarity Measurements Using Wiener-Based Indices for the Prediction of the NPY Y5 Receptor Antagonist Capacity of Benzoxazinones. *J. Chem. Inf. Model.* **2007**, *47*, 2235–2241.
(41) Urbano Cuadrado, M.; Luque Ruiz, I.; Gómez-Nieto, M. A. Refinement and Use of the Approximate Similarity in QSAR Models for Benzodiazepine Receptor Ligands. *J. Chem. Inf. Model.* **2006**, *46*, 2022–2029.
(42) Bagchia, M. C.; Maiti, B. C.; Bose, S. QSAR of Anti Tuberculosis Drugs of INH Type Using Graphical Invariants. *J. Mol. Struct. (Theochem)* **2004**, *679*, 179–186.
(43) Seydel, J. K.; Schaper, K. J.; Wempe, E.; Cordes, H. P. Mode of Action and Quantitative Structure−Activity Correlations of Tuberculostatic Drugs of the Isonicotinic Acid Hydrazide Type. *J. Med. Chem.* **1976**, *19* (4), 483–492.
(44) *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Amsterdam, The Netherlands, 1999.
(45) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley-Interscience: New York, 2000.
(46) Wiener, H. J. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17–20.
(47) Randić, M. Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.
(48) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399–404.