A Criteria To Classify Biological Activity of Benzimidazoles from a Model of Structural Similarity

Marcela Niño V.,[†] Edgar Eduardo Daza C,*[,†] and Myriam Tello[§]

Grupo de Química Teórica, Universidad Nacional de Colombia, Santa Fe de Bogotá, D.C., Colombia

Received June 30, 2000

We have classified a set of 250 benzimidazoles using a criterion of structural similarity. This criterion has led us to several clusters, which keep a close relationship between the molecules belonging to each one of them and their pharmacological activity. To study the structural similarity we have built a mathematical space where chemical structures are pictured as vectors. A set of well-chosen descriptors was used as variables. These descriptors arise from graph theoretical studies and quantum mechanical calculations. Principal components analysis was employed to find the suitable dimension for the space. Finally, cluster analysis was performed to classify the set of molecules by similarity. A Euclidean metric was used as a similarity coefficient.

1. INTRODUCTION

Chemical similarity is a key field in chemical and pharmaceutical research, although its importance is widely accepted, it is necessary to recognize that most of the current models are incomplete, ambiguous, and in some sense of local applicability. When a property is proposed as a measure of structural similarity, one of the major problems that one faces is the necessity to overlap chemical structures to be compared.

The model proposed here avoids this necessity by considering chemical structures in a different way. We do not figure out molecules as mechanical objects immersed in a three-dimensional space but as mathematical objects represented by vectors in an Euclidean space whose components describe topological and quantum features proper of their chemical nature.

We expected that these molecules will distribute in our Euclidean space according to their structural characteristics, so that, we could find neighborhoods of similar molecules. In this sense, the structural similarity could be associated with a metric for the space. For a well-defined structural space we could also expect that molecules with similar biological activity to be in the same neighborhood of structural similarity. In the first instance, a Euclidean metric was employed as a similarity measure.

We applied our model to 250 benzimidazoles used in several pathologies. Principal components analysis was employed to find the suitable dimension for the space. Finally, a cluster analysis was performed to classify by similarity the set of molecules. Thus, we get several clusters, which keep a close relationship between the molecules belonging to each class and their pharmacological activity.

We chose the benzimidazoles because they have been studied for the treatment of very diverse pathologies; they are the first drugs of choice in parasitism treatment,[1,2] a characteristic pathology of tropical areas with a strong incidence in the profiles of morbidity and mortality in the area. It also affects the agricultural sector by depressing their production and development. Also, they have been studied as antibacterians,[3,4] antifungics,[5] antivirals,[6−8] antiulcers,[9,10] anxiolitics,[11] 5-HT antagonists,[12] antihystaminics,[13] antihypertensives[14−18], anticonvulsivants,[19] antiaggregants,[20] antipsychotics,[21] antiarritmics,[22] and antineoplasitics[23] among others.

2. THE MODEL

We have defined a Euclidean space, $\boldsymbol{E}$, whose elements are chemical structures $\mathbf{Q}_k = (P_1, P_2, ..., P_m)$ characterized by a set of coordinates that are linear combinations of structural descriptors, $P_j = \Sigma d_{jk}$. These descriptors arise from graph theoretical studies and quantum mechanical calculations. We propose a partition of the space $E = \{\mathbf{C}_1 \cup \mathbf{C}_2 \cup ... \cup \mathbf{C}_3\}$, into equivalence classes defined by $\mathbf{C}_k = \{\mathbf{Q}_i |\, d(\mathbf{Q}_i\, \mathbf{Q}_k) < \epsilon\}$. If biological activity is a function of chemical structure, $A_i = f(\mathbf{Q}_i)$, then the activity of each element in the class C_k must keep a relationship: $f(\mathbf{C}_k) = A_k$, i.e., $f(\mathbf{Q}_i). \approx f(\mathbf{Q}_k) \Rightarrow A_i \approx A_k$

Building the Model. First of all, it is necessary to define the variables. As a hypothesis we outline that it is necessary to consider at least two classes of descriptors to characterize the molecular structure: those which describe the relationships between the atom and their neighbors in the molecule (structural formula) and those allowing the quantification (directly or indirectly) of the interactions between themselves.[24,25]

Graph theory[26−27] was used to support the first part of our hypothesis. Correspondingly, we calculated the graph-theoretical descriptors, based upon the molecular graph for each structure, using the algorithms implemented in CODESSA.[33] The graph-theoretical descriptors which we considered enough to grasp the topological features associated with molecular structures were the following: the Wiener index

* Corresponding author phone: (57-1) 316 5000 ext. 18324; e-mail: eedaza@ciencias.unal.edu.co.

[†] Grupo de Química Teórica.

[§] Departamento de Farmacia.

10.1021/ci000071h CCC: $20.00
Published on Web 05/08/2001

*(*W*)*; the Randić of order 0−3 (χ_n)[34] indices; the Kier and Hall indices order 0−3 ($^x\chi^v$); the Kappa indices of order 1−3 ($^x\kappa$); and the flexibility index (ϕ).[35]

For the construction of the descriptors to be used in the second part of the hypothesis, quantum theory was employed. As it is well-known, from the wave function it is possible to calculate the properties of a molecular system;[36−38] some of them could be used to quantify the interactions among the atoms in the molecule.

To obtain these descriptors, we preoptimize the molecular geometry in molecular mechanics,[39] using the force field MM+. Afterward, AM1 wave function and properties calculations were made with GAUSSIAN 94,[40] after the geometry optimization. The properties that we used as interaction descriptors were as follows: the magnitude of the inertial moment vector of the molecule, the magnitude of the dipolar moment vector,[41] the Hartree−Fock energy,[42] the energies associated to the HOMO and LUMO orbital,[43] total electronegativity (χ_{abs}), and hardness (η).[44]

Once chemical structure characterization through calculated descriptors was completed, we analyzed these variables in order to build up an appropriate workspace. The molecules and the descriptors constitute a matrix which can be considered in two ways: (i) as *n* molecules defined in a *p*-dimensional space or (ii) as *p* variables (descriptors) defined in a *n*-dimensional space.

The initial aim was to establish an appropriate mathematical space for the study of the chemical structure; therefore, we chose the second form of considering the set. Now, it was necessary to establish the most relevant descriptors in the molecular characterization. Different methods and techniques exist to do this. We used principal components analysis,[45−47] because it allows us to establish new variables—principal components (PC)—which summarize the maximum residual variability of the system.

In other words, the principal components still retain much of the descriptive power of the original data. So that, it is possible to count for the total variability with less variables; in this way we are determining a set of variables that retain the maximum variability of the original system and eliminating those ones that do not contribute with relevant information. We defined an initial space to study the chemical structure, with as many dimensions as descriptors: 20. Initially we standardized the variables to make them comparable, and by means of principal components analysis we determined the set of relevant descriptors in the study.

To define the equivalence classes and to select some molecules as representative elements to be biologically evaluated, we carried out cluster analysis.[48] In this approach, the similarity relationships between the individuals are viewed as a distance function of a set of variables. We proposed as criteria of structural similarity the Euclidean metric. If the space is well defined, i.e., if we have picked up the right descriptors, our criteria of similarity will be well behaved, and it should be reflected in the properties associated to each partition of the molecular set. Based upon this classification scheme we obtained a partition of the original set of benzimidazoles.

The next step was to study the fine structure of these subsets. Again, we applied a principal components analysis to the molecules descriptors in each subset. Thus, we defined a new local system of coordinates for each one of them. The analysis continued by means of a new clustering and the search of common biological activity or ordering in the degree of activity for the new classification of molecules.

3. RESULTS AND ANALYSIS

As it was said before, the set of molecules studied consists of 250 molecules whose possible activity like antivirals, antihelmintics, cardiotonics, and angiotensin II inhibitors (antihypertensives) was previously tested. [The values of the 20 descriptors for each one of the 250 molecules are available by solicitude.]

Definition of the Molecular Structure Space. By the study of the eigenvectors and eigenvalues (λ) generated in the principal components analysis, we established the dimension of the space to be considered and the set of relevant descriptors between those we proposed. First, we took into account the percentage of variance explained (%EV) by each principal component, PC, to choose the number of principal components to be kept as variables; variance is given by the magnitude of eigenvalues corresponding to each PC (Table 1). We have selected the eight first PCs inasmuch as they have high eigenvalues covering a 99.77% of the total variance.

To estimate the relevance of the descriptors proposed, we calculated the contribution of each one in every one of the remaining eight PCs.

Our descriptors were divided in three groups defined by their contribution to the cumulative percentage of explained variance (%CV). Thus in the range of 80−100% there are the quantum descriptors (energy, dipolar moment, inertial moment, and HOMO) and between 30 and 50% there are LUMO, total electronegativity χ_{abs}, hardness η, Wiener index *W*, the valence indices $^2\chi^v$, $^3\chi^v$, the third-order shape index $^3\kappa$, and the flexibility index ϕ. Finally, between 5 and 25% we found the following: the connectivity indices χ_0, χ_1, χ_2, χ_3, the valence indices $^0\chi^v$, $^1\chi^v$, and the shape indices $^1\kappa$, $^2\kappa$.

The last set of descriptors was excluded of the analysis because of their low relevance in definition of the variables. Note that if the number of variables is increased until reaching the 12 PCs, the percentages of explained variance that corresponds to each descriptor does not exceed 40%, except for the first-order valence index $^1\chi^v$, which becomes significant just in the 12th principal component.

Besides, after considering the meaning of each index, we saw that Randić descriptors (describing the molecular ramification without taking into account atomic identity) offered less information than the connectivity index of Kier and Hall. The latter gave a better description and discrimination of the molecular graph. In the same sense, the connectivity indices $^0\chi^v$ and $^1\chi^v$ were eliminated since they are involved in the second- and third-order connectivity index. The same thing occurred with the shape descriptors $^1\kappa$ and $^2\kappa$, inasmuch as they are involved in the index $^3\kappa$ and ϕ.

The principal components analysis led us to a new set of relevant or important indices to achieve a good description and therefore discrimination between molecular structures. After this analysis the initial 20 descriptors (seven of them of quantum origin and 13 of them graph-theoretical indices) were reduced to only 12 (five graph-theoretical and seven of quantum origin). [We must say that the inertial moment was calculated on the basis of an optimized geometry with

Table 1. Principal Components of the Original Space

	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7	PC_8	%C
				Principal Components Eigenvalues					
λ	13.50	2.77	1.52	1.09	0.49	0.36	0.12	0.09	
				Percentage of Explained Variance by Principal Components					
%EV	67.51	13.88	7.62	5.44	2.47	1.82	0.60	0.43	
%CV	67.51	81.39	89.01	94.45	96.92	98.74	99.34	99.77	
				Percentage Explained Variance of Each Descriptor in Each PC and Accumulate up to PC_8					
HF energy	0.05	2.23	49.51	2.01	25.12	2.54	16.25	1.53	99.25
dipolar M.	1.27	1.14	28.29	4.53	62.07	2.54	0.02	0.05	99.92
HOMO	0.86	0.61	4.00	73.66	0.27	0.80	0.00	0.58	80.79
LUMO	0.22	33.99	1.77	0.00	0.00	0.04	0.01	0.00	36.04
χ_{abs}	0.53	30.27	0.29	7.64	0.05	0.01	0.01	0.10	38.90
η	0.02	29.74	4.11	9.98	0.01	0.26	0.01	0.05	44.15
inertial M.	5.07	0.01	0.00	0.85	5.47	74.91	0.68	3.36	90.36
W	6.69	0.16	0.09	0.10	0.47	16.97	1.77	12.71	38.97
χ_0	7.35	0.05	0.00	0.00	0.01	0.12	0.08	4.41	12.03
χ_1	7.32	0.02	0.11	0.00	0.09	0.23	0.02	5.92	13.73
χ_2	7.28	0.00	0.13	0.00	0.11	0.31	1.57	9.82	19.22
χ_3	7.15	0.00	0.92	0.00	0.57	0.38	1.97	13.08	24.07
${}^0\chi^v$	7.35	0.07	0.00	0.10	0.14	0.25	0.42	0.01	8.34
${}^1\chi^v$	7.29	0.04	0.04	0.25	0.42	0.19	2.34	4.63	15.20
${}^2\chi^v$	7.10	0.01	0.14	0.37	0.83	0.29	12.06	11.82	32.64
${}^3\chi^v$	6.91	0.00	0.66	0.53	1.56	0.06	22.36	14.49	46.58
${}^1\kappa$	7.30	0.16	0.50	0.00	0.06	0.04	0.13	0.07	8.27
${}^2\kappa$	7.08	0.34	1.26	0.00	0.28	0.02	8.49	1.72	19.20
${}^3\kappa$	6.44	0.57	4.73	0.00	1.64	0.00	20.49	6.37	40.26
ϕ	6.70	0.56	3.42	0.00	0.81	0.00	11.29	9.23	32.03

Table 2. Principal Components of Global Structural Space

structural	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7	PC_8	%C
				Principal Components Eigenvalues					
λ	6.57	2.10	1.40	1.00	0.48	0.32	0.07	0.05	
				Percentage of Explained Variance by Each Principal Component					
%EV	54.72	17.52	11.68	8.34	3.99	2.67	0.60	0.39	
%CV[a]	54.72	72.24	83.92	92.26	96.25	98.92	99.52	99.91	
				Percentage Explained Variance of Each Descriptor for Each PC and Accumulate up to PC_8					
HF energy	0.49	19.46	25.74	4.44	24.41	7.31	17.18	0.93	99.97
dipolar M.	3.06	2.99	20.88	19.48	48.57	4.95	0.05	0.00	99.99
HOMO	1.44	2.53	28.23	45.13	0.26	1.14	0.15	0.16	79.04
LUMO	8.47	13.57	0.03	15.19	0.30	1.26	0.79	0.06	39.67
χ_{abs}	4.26	23.33	15.78	0.79	0.04	0.13	0.38	0.00	44.71
η	7.79	18.65	2.32	5.99	0.20	0.86	0.73	0.03	36.55
inertial M.	9.58	2.92	0.18	3.82	17.81	55.83	0.13	9.70	99.96
W	13.20	1.53	0.07	0.18	1.73	18.25	3.33	61.35	99.66
${}^2\chi^v$	13.46	2.44	0.40	1.98	1.68	5.12	13.60	3.03	41.72
${}^3\chi^v$	13.06	2.01	1.10	2.96	3.06	5.08	28.25	2.56	58.10
${}^3\kappa$	12.36	5.44	3.10	0.022	1.38	0.00	24.72	9.13	56.16
ϕ	12.83	5.14	2.16	0.00	0.55	0.05	10.70	13.03	44.46

respect to the quantum energy (AM1), and it is for this reason that it is included in the quantum origin descriptors, albeit, its description of the chemical structure has a topologic sense.]

This new set of descriptors was used to build a global structural space, where the principal components analysis was applied again in order to evaluate the corresponding eigenvalues and eigenvectors (Table 2), which established the self-consistence of the number of PC variables, that are involved in the maximum variability of the system and ponder the relevance of the new reduced set of descriptors.

Considering that the first eight components account for the 99.91% of the explained variance, we retained the idea of working with the eight first principal components, selected before. The descriptors kept the preponderance shown before, so that neither of them was susceptible of elimination. So, the workspace stays with eight variables, each one results from a lineal combinations of the 12 descriptors previously selected.

The first principal component (PC_1) is conformed by the topological descriptors including the inertial moment. They explain 54% of the system total variance. The PC_2 to PC_5 are formed from quantum origin descriptors, and they account for 41.5% of the variance. It is important to point out that they describe features as molecular energy, reactivity, and polarity. The PC_6, PC_7, and PC_8 cover indices that are used to illustrate the shape, size, and molecular ramification, in combination with polar features and reactivity. They involve 4% of the total variance.

In general, we found that in the definition of the structural space both types of descriptors were important, although the quantum descriptors seem to be slightly more relevant. According with these results the necessary elements to describe the molecules of this study consist of topological

Table 3. Comparison of the Structural Space[a]

space	global	branch R_A	branch R_B
%CV	99.9116	99.7110	99.8145
HF energy	12.50	12.48	12.38
dipolar M.	12.50	12.49	12.50
HOMO	9.88	9.69	9.21
LUMO	4.96	5.23	6.84
χ_{abs}	5.59	5.42	4.38
η	4.57	4.65	4.56
inertial M.	12.50	12.41	12.49
W	12.46	12.43	12.48
$^2\chi^v$	5.21	5.20	5.75
$^3\chi^v$	7.26	7.68	6.76
$^3\kappa$	7.02	5.97	7.30
ϕ	5.56	6.34	5.35

[a] Percentage of explained variance accumulated up to PC_8.

features as size, ramification and shape, and elements, which specify energetic interactions as polarity and reactivity. This result corroborates our hypothesis in the sense that the most important variables cover nearly independent topological information (the first PC) and energetic interactions (PC2−PC5).

Subsets of the Molecular Space. Once the mathematical space representing chemical structure was established, we studied the global set of molecules in terms of structural similarity. As we have proposed the Euclidean metric as a measure of similarity, it is necessary to remember that the distance between two sets, A and B, is defined as the shortest distance between elements belonging to each set, i.e., inf(d(a,b)); a ∈ A and b ∈ B.

By means of cluster analysis, we obtained two major classes, branch A (R_A) and branch B (R_B); the distance between these two sets, d(R_A,R_B), is about 1.61 units. In Figure 1, the resulting cluster can be observed. Branch A includes the majority of the molecules studied as angiotensin II receptor inhibitors;[14] all of them are closer than 1.43 units. Branch B contains the molecules that have been studied as antivirals,[6] antihelmintics,[2] and cardiotonics[49] and some studied as angiotensin II inhibitors; this set has a diameter around 1.14 units.

It is important to point out that the similarity definition in terms of neighborhoods in structural space is satisfactory. We accomplished a primary classification, which reproduces in a good way the different types of molecules proposed as the universe of work; we obtained subsets that are clearly established and each one of them is constituted by molecules generated from a common pattern structure (all of them with similar biological activities). In other words, we constructed a quantitative model adequate to quantify structural similarity. We proceed with the study of each one of the subsets independently.

Characteristics of Molecular Subset R_A. The principal components evaluated for this subset follow the same tendency as the ones of the global space. It is important to remark that the combination of descriptors reflects a major relevance in the relative size of ramification and reactivity (measured by quantum descriptors) instead of the global topology (Table 3).

The cluster obtained with the local reference system (Figure 2) exhibits the formation of three main subsets, R_A1, R_A2, and R_A3; besides there are some other molecules. The distances between these subsets, within their

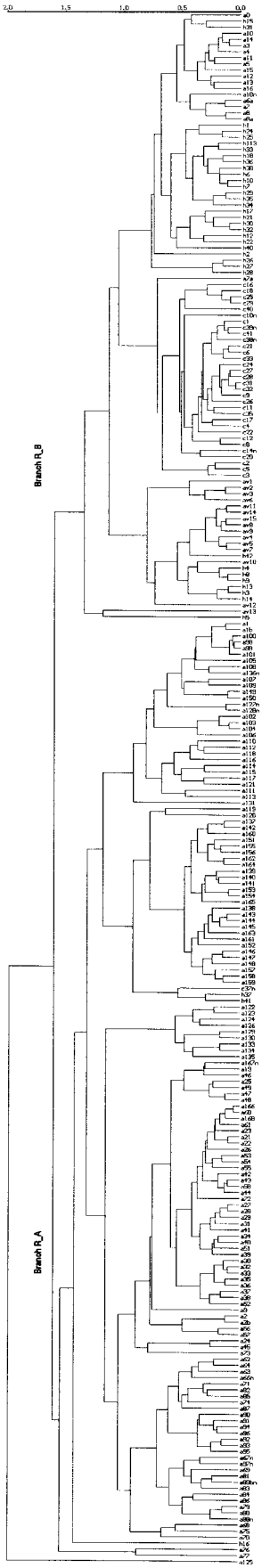

Figure 1. Global tree, using 12 PC and eight descriptors.

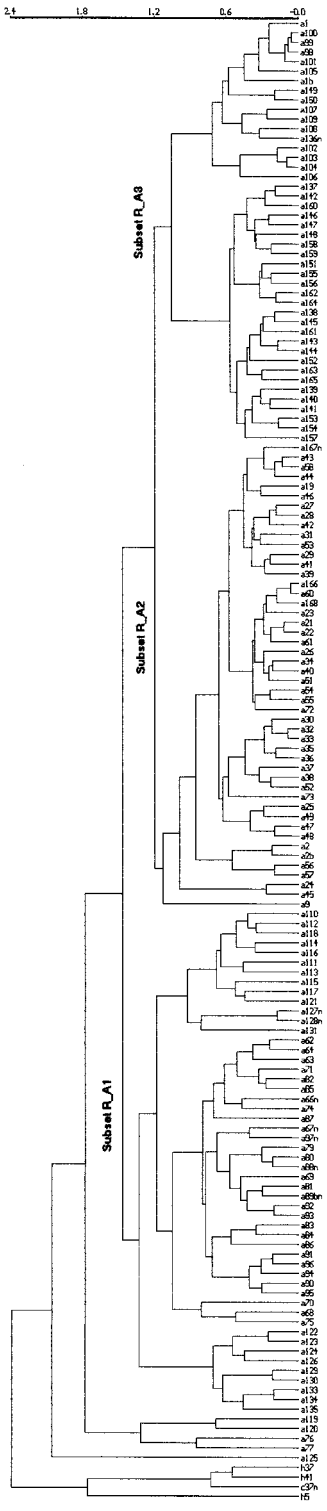

Figure 2. Tree for branch R_A.

local reference system, are as follows: d(R_A1, R_A2) ≅ 1.48, d(R_A1, R_A3) ≅ 1.48, and d(R_A2, R_A3) ≅ 1.21 The diameter of each subset is as follows: 1.34 units for R_A1, 1.14 for R_A2, and 1.06 for RA_3; these values show that we have obtained a good partition of R_A. We will come back later to the study of these subsets.

Characteristics of Molecular Subset R_B. The local reference system established for this subset is similar to the global one. Albeit, the descriptors χ_{abs} and ${}^3\chi^v$ diminish their participation, LUMO increased it. Their variables are formed by topological and quantum descriptors in equal proportion (Table 3). The PC_1 is essentially topological, while PC_2 to PC_5 are formed by quantum descriptors. Beyond this point, both types of descriptors are combined, in some cases in order to describe the differences of size and shape, and in other cases polar and energetic features combine themselves with molecular ramification indices. This suggests that when we refine the system of reference, we make clearer the differences between the type of fine ramification and the global topology. Also, indices related with atomic interaction, and in some sense with molecular reactivity, increase their importance (Table 3).

The subset R_B (Figure 3) subdivides in four subsets, R_B1, R_B2, R_B3, and R_B4. Branch R_B1 contains the whole of the molecules with antiviral activity[6] and one molecule recognized as antihelmintic (diameter ≅ 0.90). The branch R_B2 (diameter ≅ 1.03) involves only molecules studied as cardiotonics.[49] The branch R_B3 (diameter ≅ 0.92) includes only antihelmintics,[2] and the branch R_B4 (diameter ≅ 1.17) comprise molecules studied as angiotensin II receptor inhibitors and some antihelmintics.

The distances between these subsets, within their local reference system, are as follows: d(R_B1,R_B2) ≅ d(R_B1,R_B3) ≅ d(R_B1,R_B4) ≅ 1.58, d(R_B2, R_B3) ≅ d(R_B2, R_B4) ≅ 1.47, and d(R_B3, R_B4) ≅ 1.25.

The local system of reference that we have established has the capability to classify molecules by their biological activity. The quality of this classification may be inferred comparing the diameter of each subset and distances between each of them. We repeat the procedure, to establish new local systems of reference for each subset of R_A and R_B and thus to make more sensitive the subspaces in the definition of molecular similarity.

Molecular Subsets of Branch R_A. As we mentioned earlier, this branch contains molecules designed and tested as angiotensin II receptor inhibitors. We continue with the branches in which this subset is subdivided.

- Molecular subset R_A1a: In this subset (Figure 4) molecules studied by Ries[17] appear. Based on Losartan, and with a systematic variation of different substituents in different positions at the benzimidazole ring, they determine that with the substitution of the 6-position acylamino groups, highly active angiotensin II antagonists are obtained. Our clustering classifies them in two sets, one of high activity and the other one of low activity.

- Molecular subset R_A1b: In this subset there are a group of molecules synthesized and tested by Kubo et al.,[15] designed to improve the oral biodisponibility of CV-11974 (a43) and CV-11194 (a168)—strong angiotensin II antagonists. The cluster (Figure 5) indicates two subclasses of molecules, which reflect the substitution type, made over

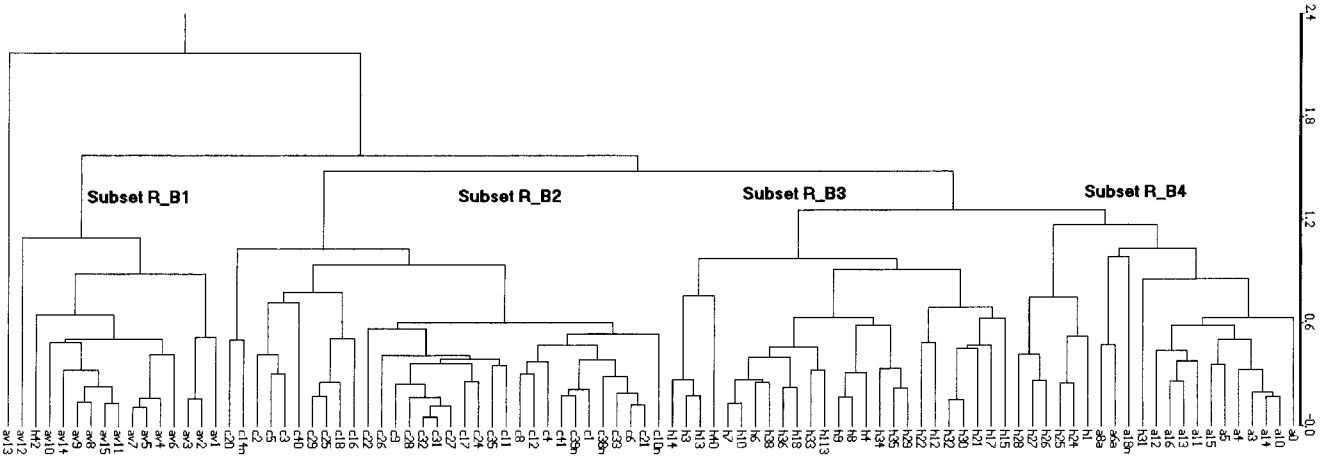

Figure 3. Tree for branch R_B.

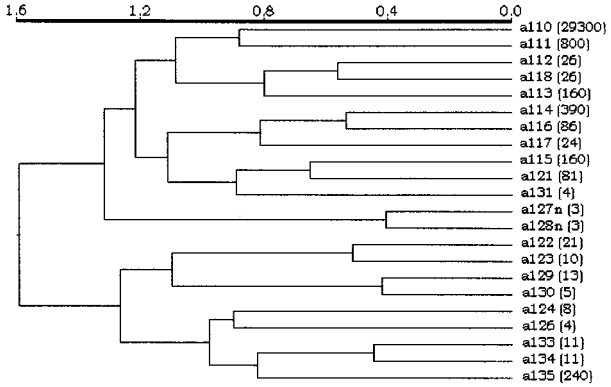

Figure 4. Tree for subset R_A1a. Values in parentheses correspond to IC_{50} for specific binding of [^{125}I] angiotensin II to rat lung preparation (see ref 17).

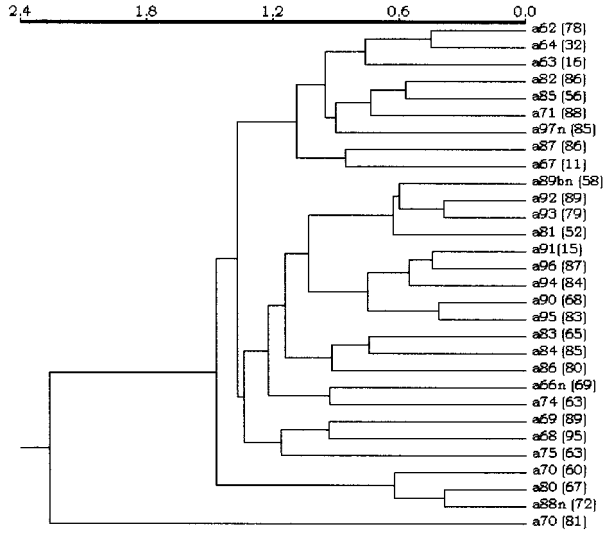

Figure 5. Tree for subset R_A1b. Values in parentheses correspond to percent inhibition of the angiotensin II (100ng/kg iv) induced pressure response to each time after administration of the test compounds in conscious male Sprage-Dawley rats (ref 15).

the base nucleus, but a relationship with the degree of activity is not present.

• Molecular subset R_A2: In this subset there are the CV-11194 (a168) analogous molecules synthesized and tested by Kubo[14] et al., to study the structure−activity relationships of this molecule. The cluster obtained (Figure 6) reflects the substitution in the nucleus.

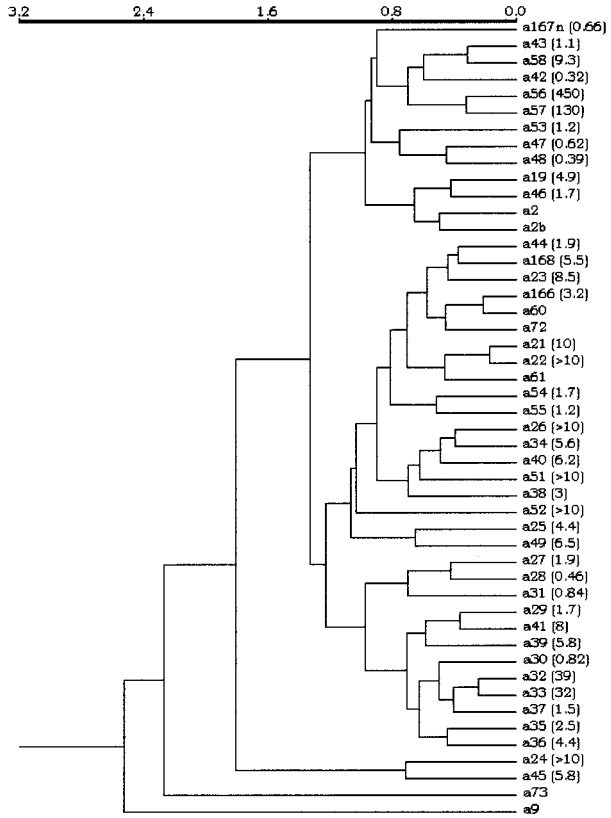

Figure 6. Tree for subset R_A2. Values in parentheses correspond to inhibition of specific binding of [^{125}I] angiotensin II (0.2 nM) to bovine adrenal cortex. Values are expressed as IC_{50} ($\times 10^{-7}$) (see ref 14).

• Molecular subset R_A3: The molecules, which constitute this subset, differ in their shape and ramification but maintain a similarity in their polarity and reactivity. The cluster (Figure 7) exhibits two subsets, R_A3a and R_A3b. Like the two previous subsets, there is no relation with biological response.

• Molecular subset R_A3a: The molecules of this subset (Figure 8) are part of the study of Ries,[17] as the subset R_A1a. In both cases, the efforts were aimed to characterize the nature of the bounding of the benzimidazole to the angiotensin II receptor, AT_1. The most remarkable differences in this subset are the topological features of the molecules.

• Molecular subset R_A3b: This subset contains molecules of Kohara[16] studies, which are based upon CV-11974 (a43), TCV-116 (a96), and TAK-536 (a165) molecules, and through

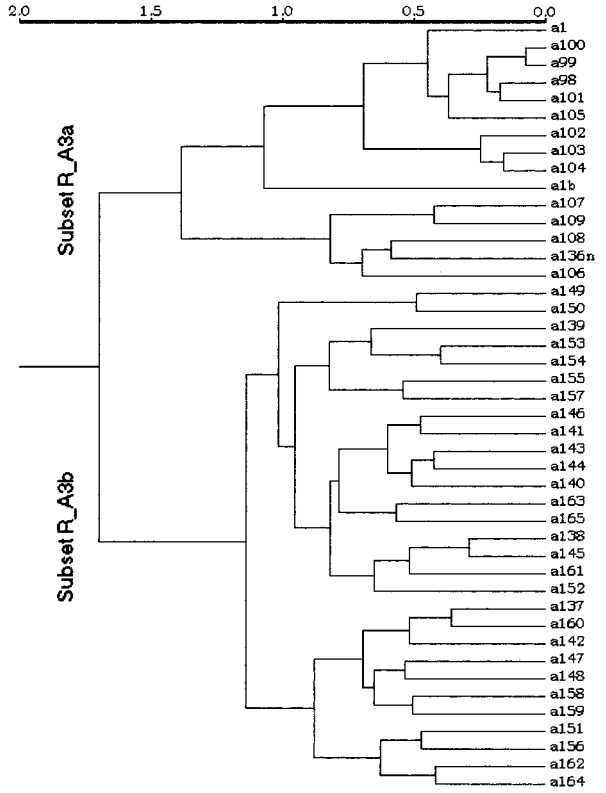

Figure 7. Tree for subset R_A3.

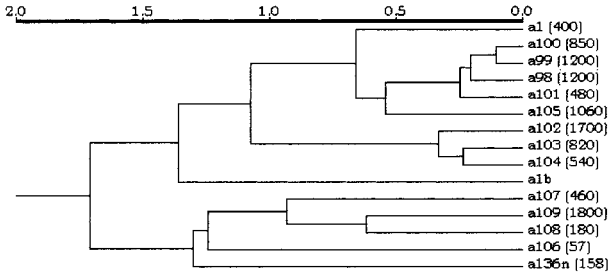

Figure 8. Tree for subset R_A3a. Values in parentheses correspond to inhibition of specific binding of [^{125}I] angiotensin II (0.2 nM) to bovine adrenal cortex. Values are expressed as IC_{50} ($\times 10^{-7}$) (see ref 17).

bioisosteric substitutions on this nucleus are searched for analogous with comparable activity, evading the synthetic, metabolic, and chemical disadvantages of the tetrazol. The cluster obtained (Figure 9) separates molecules in accordance with the substitution type made over the base nuclei.

Finally, we compared subsets of the Branch R_A. The conclusion is that there are several indices having a high and constant incidence in the definition of the subspaces: energy, dipolar, and inertial moments. The remaining indices exhibit important variation in each subset. Indices of quantum origin possess the major discriminatory power, as they explain the principal percentage of variance in each case. In contrast, descriptors varying from group to group reflect the internal variations of each one of the subsets (Table 4).

The subset R_A1 is divided into the branches R_A1a and R_A1b. The descriptors HOMO, Wiener, and ϕ increase in the subsets R_A1a and R_A1b, while the χ_{ABS}, $^2\chi^v$, and $^3\chi^v$ indices diminished their contributions. Results that differ from the situation in R_A1. The $^3\kappa$ and η indices remained constant, in medium and low levels, respectively. This indicates that the separation of each subset is based on

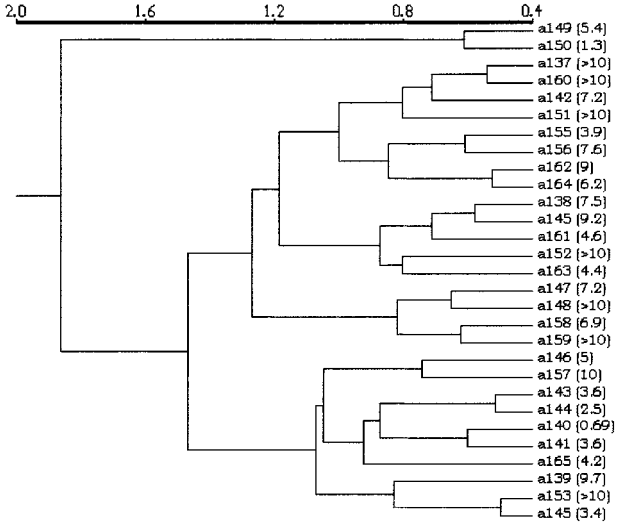

Figure 9. Tree for subset R_A3b. Values in parentheses correspond to inhibition of specific binding of [^{125}I] angiotensin II (0.2 nM) to bovine adrenal cortex. Values are expressed as IC_{50} ($\times 10^{-7}$) (see ref 16).

Table 4. Comparison of the Structural Subspaces of the Branch A

space	R_A	R_A1	R_A1a	R_A1b	R_A2	R_A3	R_A3a	R_A3b
%CV	99.71	98.96	99.85	99.30	99.53	99.76	99.98	99.94
HF energy	12.48	11.84	12.18	12.49	12.00	12.38	12.47	12.45
dipolar M.	12.49	12.50	12.50	12.39	12.48	12.41	12.48	12.48
HOMO	9.69	6.45	9.96	10.49	7.00	8.80	6.58	7.78
LUMO	5.23	5.18	4.65	5.07	9.01	8.37	10.04	8.63
χ_{abs}	5.42	7.53	5.86	4.16	4.55	3.71	4.39	4.32
η	4.65	5.51	4.47	4.31	4.37	3.79	3.95	4.20
inertial M.	12.41	12.13	12.48	9.30	9.17	11.47	11.36	12.09
W	12.43	4.66	11.61	9.32	11.94	11.48	4.64	6.52
$^2\chi^v$	5.20	12.36	5.62	9.50	6.63	6.37	11.38	10.39
$^3\chi^v$	7.68	11.19	7.22	8.68	8.80	7.03	11.26	11.70
$^3\kappa$	5.97	5.80	5.30	5.59	4.64	6.42	4.15	6.01
ϕ	6.34	4.86	8.16	8.69	9.41	7.77	7.30	3.43

similarity of features such as polarizability, reactivity, and shape and on energy, polarity, size, and ramification differences.

In branch R_A2 the descriptors contribution is established at two levels. The first level is between 9 and 12% (energy, dipolar moment, LUMO, inertial moment, Wiener, $^3\chi^v$, and ϕ), and the second one is from 4 to 7% (χ_{abs}, η, $^2\chi^v$, and $^3\kappa$). Analogously, to the previous subset, this is a form based on features of structural similarity such as polarizability and molecular reactivity. The resulting discrimination depends on shape, size, and polarity features.

Finally, with respect to branch R_A3 and its subsets, the classification is based on the similarity of features encoded by the indices χ_{abs}, η, W, and $^3\kappa$. The HOMO, LUMO, $^2\chi^v$, and $^3\chi^v$ descriptors have an important incidence at the core of each subset.

In general, the features in which the molecular similarity is established for the definition of each subset are polarizability and reactivity, and at the core of each set the features are polarity, ramification, and shape (Table 4).

Molecular Subsets of Branch R_B. A new local frame of reference was set for each one of the subsets in this branch.

The Subset R_B1. Molecules in this subset have been studied as antiviral drugs. The cluster we obtained for this branch, using a new local frame, reflects the type of substitution on the benzimidazol ring and it divides into two

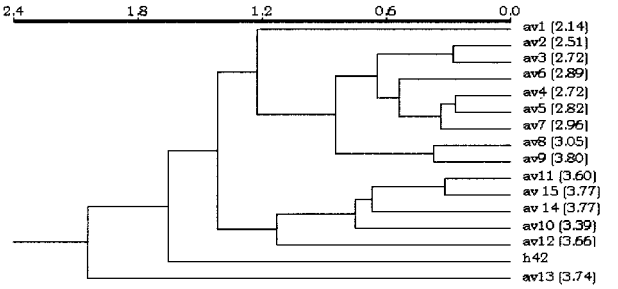

Figure 10. Tree for subset R_B1. Values in parentheses correspond to antiviral values expressed as the logarithm of concentration to give 75% inhibition of multiplication log(1/c) (see ref 6).

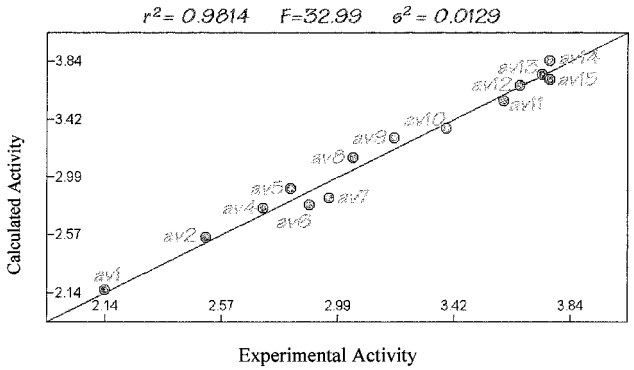

Figure 11. Correlation between experimental and calculated activity for antiviral subset R_B1.

subsets. These subsets are closely related to the intensity of the biological response of their molecules (Figure 10). Besides, there are two more molecules showing a different behavior, i.e., they are outside of the two main branches, already seen in the global tree.

Additionally, we evaluated the relationship between the principal components and the activity through multivariate regression analysis, to obtain a quantitative structure−activity relationship. An equation with $r = 0.99066$, $r^2 = 0.98141$, and $F = 32.99027$ was obtained. Experimental and calculated biological activities were plotted to see their correlation; the results are satisfactory (Figure 11). We also tested the correlation between each PC with biological activities; the best equation was achieved with the PC_1 ($r = 0.95960$, $r^2 = 0.92083$, and $F = 151.2111$), which its characteristic feature is that it is a mixture of descriptors coming from both origins, graph theoretical and quantum mechanical.

The h42 molecule seemed to have antiviral activity, based on the multivariate regression terms. Although in our work, this molecule appears as a member of this cluster it was experimentally studied as antihelmintic. The resulting data showed a potential antiviral activity for this molecule. In addition, considering that Kier and Hall[6] pointed out that antiviral activity depends on ramification of the substituent in the 2-position of the benzimidazole ring and that h42 molecule has in this position a tiazolil radical, the experimental test of this molecule will be of most interest.

Subset R_B2. Molecules in this subset have been studied as cardiotonic drugs. The cluster for this subset (Figure 12) presents the formation of three molecular branches defined by their ramification and energetic differences. The most external corresponds to the majority of molecules whose activity cannot be determined. This result can be related with (i) features making difficult their binding to the receptor or

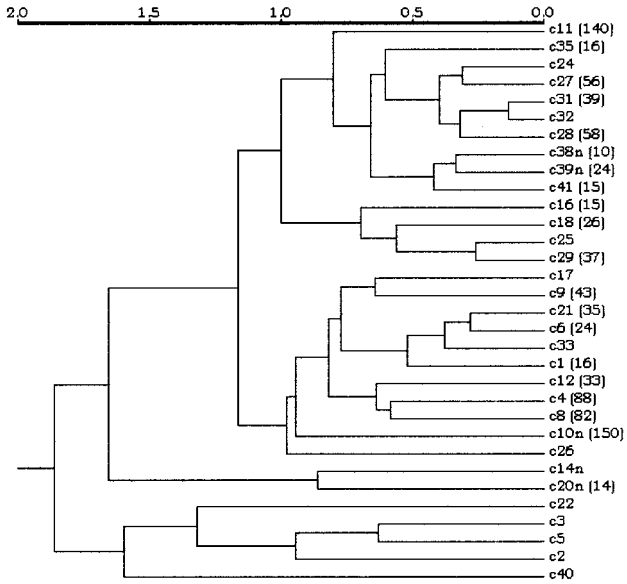

Figure 12. Tree for subset R_B2. Values in parentheses correspond to EC_{50} μM, isotropic activity in isolated left atria (see ref 49).

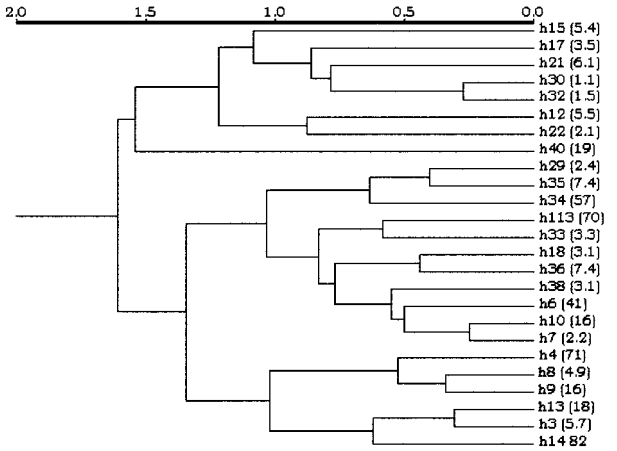

Figure 13. Tree for subset R_B3. Values in parentheses correspond to IC_{50} μM to inhibition of tubuline polymerization in sheep brain (see ref 2).

(ii) the lack of the binding requirements. In the other two clusters, we saw that molecules which are more distant to the principal branches of their respective tree present a lower cardiotonic activity.

Subset R_B3. The molecules of this subset are antihelmintics, and they differ in ramification and reactivity and present uniformity of size and shape. The cluster (Figure 13) reflects the type of substitution on the benzimidazol nucleus. There are two clusters of molecules, one with alkyl substituents and the other with highly polar substituents.

Subset R_B4. The previously discussed subsets consist of molecules with a unique activity. The R_B4 subset (Figure 14) includes molecules with recognized antihelmintic activity and some others which have been studied as angiotensin II inhibitors (synthesized by conformational analogy). Due to the fact that their biological activity is sufficiently high, we suggest evaluating both clusters of molecules interchanging the target; those ones studied as antihelmintics must be tested as angiotensin II inhibitors and vice versa.

In the branch R_B subsets, the most discriminant descriptors were energy and dipolar moment. These two descriptors with electronegativity remained constant in the definition of

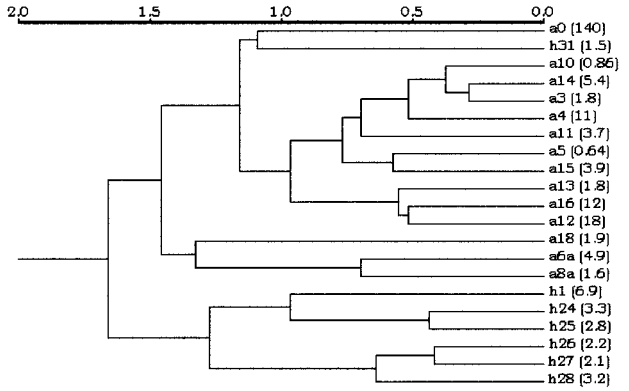

Figure 14. Tree for subset R_B4. Values in parentheses corresponds to IC_{50} μM inhibition of angiotensin II receptor binding in a normotensive rat model (see ref 18). For molecules hXX see Figure 13.

Table 5. Comparison of the Structural Subspaces of the Branch B

space	R_B	R_B1	R_B2	R_B3	R_B4
%CV	99.82	99.97	99.75	99.71	99.97
HF energy	12.38	11.71	12.15	12.28	12.34
dipolar M.	12.50	12.47	12.13	12.22	12.49
HOMO	9.21	6.31	8.02	8.43	7.07
LUMO	6.84	10.23	7.47	7.63	8.41
χ_{abs}	4.38	3.96	4.53	3.99	4.64
η	4.56	3.53	4.58	4.09	4.68
inertial M.	12.49	11.87	12.47	6.50	12.35
W	12.48	4.68	10.64	7.84	12.43
$^2\chi^v$	5.75	11.02	7.57	10.98	5.65
$^3\chi^v$	6.76	10.35	8.56	10.13	6.67
$^3\kappa$	7.30	7.58	8.60	6.94	6.39
ϕ	5.35	6.29	3.28	8.95	6.90

the different local frames. The remaining descriptors vary their influence sensitively when defining the PCs. LUMO acquired importance, while HOMO, inertial moment, and Wiener diminished their contribution when we set the local systems of reference. The behavior changes notoriously in every subset (Table 5).

In branch R_B1, (antivirals) the descriptors related with molecular reactivity—are homogeneous in their contribution, while the topological indices of ramification and shape increase their contributions to the principal components. In subset R_B2 (cardiotonics) there is a big difference between the importance played by energy, polarity, shape, and ramification descriptors. In contrast, HOMO, electronegativity, and hardness contribute in a similar way.

For subset R_B3 (antihelmintics) the contribution of each descriptor exhibits a behavior analogous to the one of the previous cases, but differences in size and shape diminish slightly their contribution in some of them. Simultaneously, ramification increases its discriminant character. The local frame for R_B4 has equivalent contributions of the two kinds of descriptors we proposed to employ. It emphasizes our finding about the importance of the energy, dipolar moment, and inertial moment indices. Features such as polarizability, reactivity, ramification, and shape participate in a similar way in the formation of PCs, which reflect the molecular diversity in this subset.

Summarizing we can say that in branch R_A we clustered the molecules into subsets which coincided with the different groups proposed and tested by diverse authors[14−18] who studied these molecules attempting to recognize the structural determinant keys of the inhibition of the angiotensin II. The subsets that we found reflect the systematic substitution of different molecular groups at the same nucleus, a strategy which is followed in the mentioned studies. In branch B, the subsets we found have a concordance with common structural features, which define the different activities, antihelmintics,[2], antivirals,[6] and cardiotonics.[49] This shows that our model achieves to separate the distinct sort of molecules, on the basis of the structural similarity, which is narrowly related with the biological activity.

4. CONCLUSIONS

We developed a quantitative model capable of classifying by structural similarity a set of 250 benzimidazoles. The classification obtained is sensitive to the type and position of the substitution on the benzimidazol ring. In certain cases, the classification led us to subsets related to the type or grade of pharmacological activity.

To establish a criterion of structural similarity it is necessary to characterize the molecular structure by quantum and topological properties. The first involves polarity and molecular reactivity, and the second one engages shape, size, and ramification; all of them are essential in the definition of variables to encode molecular structure.

The principal components analysis allowed us to select an appropriate set of variables, which conveniently describes the molecular diversity of the benzimidazoles under study. So that, it was possible to classify the set of molecules using only 12 of the 20 descriptors originally proposed and eight PCs, formed as a linear combination of those descriptors.

The quantum origin descriptors were fundamental in the conformation of the PCs and remained as necessary variables to characterize the molecules in each system of reference. In other words, they are essential for their high discriminant value. When defining the subsets by similarity, the topological descriptors increased their importance in the definition of local frames and seemed to be crucial when differentiating molecules within each subset.

The characterization of the molecular structure as an abstract mathematical object in a n-dimensional space defined by a select number of variables, instead of conceiving it as a geometrical object in a tridimensional space, was successful, as we obtained a criterion for quantifying molecular similarity based upon the Euclidean distance and through it a classification of the molecules.

REFERENCES AND NOTES

(1) PAHO. Communicable Diseases. Control of Intestinal Parasitosis; http://165.158.1.110/english/hcp/hctpep01.htm 1999.

(2) Lacey, E. The role of the cytoeskeletal protein, tubulin, in the mode of action and mechanism of drug resistance to benzimidazoles. *Int. J. Parasitol.* **1988**, *18*, 885−890.

(3) Kus, C.; Goker, H.; Ayhan, G.; Ertyan, R.; Altanlar, N.; Akin, A. Synthesis and antimicrobial activity of some new piperidinil benzimidazoles. *Farmaco* **1996**, *51*, 413−417.

(4) Goker, H.; Tebrizli, E.; Abbasoglu, U. Synthesis of 1,2, 5,(6)-trisubstituted benzimidazoles and evaluation of their antimicrobial activities. *Farmaco* **1996**, *51*, 53−58.

(5) Going, C. J.; Mayer, V. W. Induction of chromosome loss in *Saccharomyces cerevisiae* strain D61.M by selected benzimidazole compounds. *Mut. Res.* **1995**, *343*, 185−199.

(6) Hall, L.; Kier, L. Molecular Connectivity and substructure analysis. *J. Pharm. Sci.* **1978**, *67*, 1743−1747.

(7) Zou, R.; Ayres, K.; Drach J.; Townsend, L. Synthesis and antiviral evaluation of certain disubstituted benzimidazole ribonucleosides. *J. Med. Chem.* **1996**, *39*, 3477−3488.
(8) Chimirri, A.; Grasso, S.; Molica, C.; Monforte, A.; Monforte, P.; Zappala, M.; Scopelliti, R. Anti-HIV agents. IV. Synthesis and in vitro anti-HIV activity of novel 1-(2,6-difluorophenyl)-1H, 3H-thizolo-(3,4-a)benzimidazoles. *Farmaco* **51**, **1996**, *4*, 279.
(9) Terashima, K.; Muraoka, O.; Ono, M. Studies on antiulcer agents: III. Plausible mechanism of antisecretory action of ethyl 2-((1H-benzimidazol-2-yl)sulfinylmethyl)-4-(dimethylamino)-5-pyrimidinecarboxylate, an H+/K+-ATPase inhibitor, based on its reaction with thiols. *Chem. Pharm. Bull.* **1995**, *43*, 1985−1991.
(10) Cheon, H. G.; Yum, E. K.; Kim, S. S. Effects of newly synthesized benzimidazole derivatives on gastric H+/K+ATPase. *Arch. Pharm. Res.* **1996**, *19*, 126−131.
(11) Maryanoff, B.; McComsey, D.; Ho, W. Shank, R. P.; Dubinsky, B. Potential anxiolytic agents. II. Improvement of oral efficacy for the pyrido(1,2-a)benzimidazole (PBI) class of GABA-A receptor modulators. *Bioorg Med. Chem. Lett.* **1996**, *6*(3*)*, 333.
(12) Ohta, M.; Suzuki, T.; Nagashima, S.; Tokunaga, T.; Miyata, K.; Mase, T. Novel 5-hydroxytryptamine (5-HT-3) receptor antagonists. IV. Synthesys and pharmacological evaluation of the oxidation products of (−)-(R)−((1-methyl-1H-indol-3yl)carbonyl)-4,5,6,7-tetrahydro-1H-benzimidazole hydrochloride (YM060: Ramosetron). *Chem. Pharm. Bull.* **1996**, *44*, 1717−1722.
(13) Bordi, F.; Mor, M.; Morini, G.; Plazzi, P. B.; Caretta, A. *Farmaco* **1994**, *49*, 153.
(14) Kubo, K.; Kohara, Y.; Imamiya, E.; Sugiura, Y.; Inada, Y.; Furukawa, Y.; Nishikawa, K.; Naka, T. Nonpeptide Angiotensin II receptor antagonists. Synthesis and biological activity of Benzimidazolecarboxylic acids. *J. Med. Chem.* **1993**, *36*, 2182−2195.
(15) Kubo, K.; Kohara, Y.; Yoshimura, Y.; Inada, Y.; Shibouta, Y.; Furukawa, Y.; Kato, T.; Nishikawa, K.; Naka, T. Nonpeptide Angiotensin II receptor antagonists. Synthesis and biological activity of potential prodrugs of Benzimidazole-7-carboxylic acids. *J. Med. Chem.* **1993**, *36*, 2343−2349.
(16) Kohara, Y.; Kubo, K.; Imamiya, E.; Wada, T.; Inada, Y.; Naka, T. Synthesis and Angiotensin II receptor antagonists activities of benzimidazole derivatives bearing acidic heterocycles as novel tetrazole bioisosteres. *J. Med. Chem.* **1996**, *39*, 5228−5235.
(17) Ries, U.; Mihm, G.; Narr, B.; Hasselbach, K.; Wittneben, H.; Entzeroth, M.; van Meel, J.; Wienen, W.; Hauel, N. 6-Substituted benzimidazoles as a new nonpeptide Angiotensin II receptor antagonists: synthesis, biological activity, and structure-activity relationships. *J. Med. Chem.* **1993**, *36*, 4040−4051.
(18) Thomas, A.; Allott, C.; Gibson, K.; Major, J.; Masek, B.; Oldham, A.; Ratcliffe, H.; Roberts, D.; Russell, S.; Thomason, D. New nopeptide Angotensin II. 1. Synthesis, biological properties, and structure−activity relationships of 2-alkyl benzimidazole derivatives. *J. Med. Chem.* **1992**, *35*, 877−885.
(19) Isikdag, I.; Ucucu, U.; Ersan, S. Studies on the syntjesis of some 2-(1'-3'-benzodioxol-5'-yl)benzimidazole, benzothiazole, and benzoxazole derivatives. *Gaz. Univ. Eczac. Fak. Derg.* **1991**, *8*, 17−22.
(20) Sergeeva, N.; Evstigneeva, R.; Rudko, I..; Kubatiev, A.; Shorshnev, S. Synthesis and antiaggregant activity of N-substituted imidazole and benzimidazole derivatives. *Khim. Farm. Zhur.* **1995**, *29*, 13−15.
(21) Norman, M. H.; Navas, F., III; Thomson, J. B.; Rigdon, G. C. Synthesis and evaluation of heterocyclic carboxamides as potential antypsychotic agents. *J. Med. Chem.* **1996**, *39*, 4692−4703.
(22) Anisimova, V.; Spasov, A.; Bocharova, I.; Ostrovskii, O.; Panchenko, T.; Dudchemko, G. Synthesis and pharmacological activity of 4-(2-dialkylaminoethyl)pyrrolo(1,2-a)benzimidazole salts. *Khim. Farm. Zhur.* **1996**, *30*, 20−25.
(23) Kim, J. S.; Gatto, B.; Yu, C.; Liu, A.; Liu, L. F.; Lavoie, E. J. Substituted 2.5'-Bi-1H-benzimidazoles: Topoisomerase I inhibition and cytotoxicity. *J. Med. Chem.* **1996**, *39*, 992−998.
(24) Villaveces, J. L.; Daza, E. E. The Concept of Chemical Structure. In *Concepts in Chemistry*; Rouvray D., Ed.; J. Wiley and Sons: New York, 1997.
(25) Bermúdez, C. I.; Daza, E. E.; Andrade L. E. Characterization and Comparison of Escherichia coli Tranfer-RNAs by Graph Theory based on Secondary Structure. *J. Theor. Biol.* **1999**, *197*, 193−205.
(26) Harary, H. *Graph Theory*; Addison-Wesley: Reading, MA, 1969.
(27) Wilson, R. J. *Introduction to Graph Theory*; Academic Press: New York, 1972.
(28) Balaban, A. Applications of graph theory in chemistry. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 334−343.
(29) Balaban, A.; Rouvray, D. In *Applications of Graph Theory*; Wilson R. J., Ed.; Academic Press: London, 1979.
(30) Balaban, A. *Chemical Applications of Graph Theory*; Academic Press: London, 1979.
(31) Trinajstić, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1983.
(32) Milne, G. W. A. *Mathematics Basis Chem.* **1997**, *37*, 639−644.
(33) CODESSA, Comprehensive Descriptors for Structural and Statistical Analysis V2.13; Semichem 7128 Sunmit, Shawnee, KS, 1995.
(34) Randić, M. On characterization of molecular branching. *J. Chem. Am. Soc.* **1975**, *97*, 6609−6615.
(35) Hall, L.; Kier, L. The Molecular Connectivity Chi Index and kappa Shape Indexes in Structure−Property Modeling. In *Reviews in Computational Chemistry*; Lipkowitz, K., Boyd, D., Eds.; VCH: New York, 1991; Vol. 2.
(36) Levine; I. N. *Quantum Chemistry*; Prentice Hall: NJ, 1991.
(37) McWeeney, R. *Methods of Molecular Quantum Mechanics*; Academic Press: New York, 1989.
(38) Bersurker, I. B.; Dimoglo, A. S. The Electron-Topological Approach To The Qsar Problem. In *Reviews In Computational Chemistry*; Lipkovitz, K. P., Boyd, D. B., Eds.; VCH: New York, 1991.
(39) Bowen, J. P.; Allinger, N. *Molecular Mechanics: The art and Science of Parametrization*. In *Reviews in Computational Chemistry;* Lipkowitz, K. Y., Boys, D., Eds.; 1991; Vol. 2.
(40) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Gill, P. M. W.; Johnson, B. G.; Robb, M. A.; Cheeseman, J. R.; Keith, T.; Petersson, G. A.; Montgomery, J. A.; Raghavachari, K.; Al-Laham, M. A.; Zakrzewski, V. G.; Ortiz, J. V.; Foresman, J. B.; Cioslowski, J.; Stefanov, B. B.; Nanayakkara, A.; Challacombe, M.; Peng, C. Y.; Ayala, P. Y.; Chen, W.; Wong, M. W.; Andres, J. L.; Replogle, E. S.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Binkley, J. S.; Defrees, D. J.; Baker, J.; Stewart, J. P.; Head-Gordon, M.; Gonzalez, C.; Pople, J. A. *Gaussian 94*, revision D4; Gaussian, Inc.: Pittsburgh, PA, 1995.
(41) Hinchliffe, A. *Computational Quantum Chemistry*; John Wiley: 1989.
(42) Bodor, N.; Gabanyi, Z.; Wong, C. A new method for the estimation of partition coefficient. *J. Am. Chem. Soc.* **1989**, *111*, 3783−3786.
(43) Karelson, M.; Lobanov, V.; Katritzky, A. Quantum-chemical descriptors in QSAR-QSPR studies. *Chem. Rev.* **1996**, *96*, 1027−1043, and cited references.
(44) Pearson, R. The elctronic chemical potential and chemical hardness. *J. Mol. Struct. (THEOCHEM)* **1992**, *255*, 261−270.
(45) Jurs, P. C. Chemometrics and Multivariate in Analytical Chemistry. In *Reviews in Computational Chemistry*; Lipkowitz, K., Boyd, D., Eds.; VCH: New York, 1990; Vol. 1.
(46) Brereton, R. *Chemometrics;* Ellis Horwood: New York, 1990.
(47) Manly, B. *Multivariate Statistical Methods*; Chapman and Hall: USA: 1986.
(48) Brown, R.; Martin, Y. Use of structure−activity data to compare structure-based clustering methods and descriptors for use in. *J. Chem. Comput. Sci.* **1996**, *36*, 572.
(49) Güngör, T.; Fouquet, A.; Teulon, J. M.; Provost, D.; Cazes, M.; Cloarec, A. Cardiotonic Agents. Synthesis and cardiovascular properties of novel 2-Arilbenzimidazoles and Azabenzimidazoles *J. Med. Chem.* **1992**, *35*, 4455−4463.

CI000071H