Refinement and Use of the Approximate Similarity in QSAR Models for Benzodiazepine Receptor Ligands

Manuel Urbano Cuadrado,[†] Irene Luque Ruiz,*[,‡] and Miguel Ángel Gómez-Nieto[‡]

Institute of Chemical Research of Catalonia ICIQ, Avinguda Països Catalans 16, E-43007 Tarragona, Spain, and Department of Computing and Numerical Analysis, University of Córdoba, Campus Universitario de Rabanales, Albert Einstein Building, E-14071 Córdoba, Spain

Received May 28, 2006

Several considerations for refining the approximate similarity measurements have been introduced in this paper: the use of topological invariants for the calculation of similarity indexes and the development of new similarity correction processes. The quality of the new similarity measurements obtained with the proposed methods has permitted the development of fast, cheap, and simple quantitative structure−activity relationship models for the prediction of biological activities of nonbenzodiazepine γ-aminobutyric $acid_A$/benzodiazepine receptor ligands (58 compounds). Internal and external validations were carried out for the approximate similarity matrices computed using different approaches. Satisfactory results which compare reasonably well with a 3D approach were obtained: $Q^2 = 0.65$ and standard error in cross validation SECV $= 0.83$ for the training stage; $r = 0.79$ and error in external prediction $= 0.82$ for the test step. In addition, the method proposed was compared with other topological approaches based on constitutional similarity and on fingerprints. Satisfactory results were obtained.

1. INTRODUCTION

The identification, study, and use of quantitative structure−activity relationship (QSAR) approaches are important stages involved in the development of rational approaches in drug design.[1,2] In 1969, Hansch[3] demonstrated the possibility of establishing mathematical functions which correlate chemical structures with biological activities of compounds. Since then, many scientists have been working on related topics—namely, obtaining finer descriptors[4−6] (pieces of chemical information employed as predictive and modeling variables), the development of new chemometric techniques and validation strategies,[7,8] and so forth—aimed at providing pharmaceuticals with powerful tools which guide the design of new drugs. Thus, a large number of QSAR models dealing with many chemical families have been developed and summarized in the bibliography.

The validation of QSAR models[9,10] must be carried out in order to evaluate the predictive ability of equations according to accuracy, robustness, costs (resources and speed), and reversibility (simple coming-back process from equations to chemical structures). If these four characteristics are placed on the vertices of a tetrahedron, an ideal QSAR model would correspond with the center point. But taking into account the *fitness for purpose*, which characterizes technology transference, unbalanced or distortional tetrahedrons are many times optimal tools in QSAR studies. Thus, specific characteristics of models can be pursued, for example, the performance of database screening methods aimed at discriminating between compounds with low and high activities, the reversibility of 3D QSAR methods employed for the modular design of new drugs, and so forth.

Models based on topological descriptors or graph isomorphism[11−15] (computed using 2D structures representing chemical graphs) are often considered as quick methods when they are compared with other 3D approaches. The latter requires the searching and optimization of 3D conformations, which are complex and time-consuming processes when bulky and flexible molecules compose the data set.

Moreover, 3D approaches—most of them based on comparative molecular field analysis (CoMFA) and CoMFA-like methods,[16,17] which represent exactly steric and electrostatic feels of molecules by means of 3D grid interaction fields—demand the knowledge of common alignment or binding mode in order to obtain reproducible predictive spaces. Thus, despite obtaining many times better correlations with 3D approaches and designing tools, the commented shortcomings would be taken into account if simple and fast methods are pursued. Therefore, trials focused on improving the predictive ability of topological QSAR models should be carried out[18−20] with the aim of obtaining rapid and accurate predictive tools.

Regarding isomorphism calculation, we have recently developed a new similarity concept, the *approximate similarity* (AS) measurement,[15] based on computing similarities and distances of common and noncommon subgraphs, respectively.

AS values—an N by N matrix was employed as the predictive space—correlate reasonably well with steroid structures regarding other 3D approaches, and besides, simplicity is ensured by the fact of employing 2D structures. Good correlations were obtained by means of using differences among nonisomorphic substructures for correcting constitutional similarity—classical similarity values (based

* Corresponding author phone: +34-957-21-2082; fax: +34-957-21-8630; e-mail: ma1lurui@uco.es.
[†] Institute of Chemical Research of Catalonia ICIQ.
[‡] University of Córdoba.

10.1021/ci600216h CCC: $33.50
Published on Web 08/29/2006

on the consideration of the number of nodes and edges) do not achieve the predictive ability shown by 3D approaches.

The applicability of AS measurements to more flexible compounds than the rigid steroidal nucleus must be carried out in order to validate this methodology. In this paper, the refinement of AS measurements is proposed, aimed at maintaining the high predictive ability. For this purpose, AS values take into account similarity calculation based on topological invariants values instead of the use of constitutional values (number of nodes and edges of molecular graphs). In addition, new contributions of the factor which weights the influence of nonisomorphic distances in similarity are studied. A factor provided with chemical information which is not empirical and depending on the pair of structures to be compared is pursued. Thus, its optimization could be carried out automatically.

The new AS method proposed in this paper has been applied to the development of rapid QSAR models for predicting biological activities of benzodiazepine receptor ligands with nonbenzodiazepine structures. The recently developed 3D QSAR for these compounds[21] and their higher structural diversity have been, among others, the main reasons for this selection. The γ-aminobutyric $acid_A$/benzodiazepine receptor ($GABA_A$/BzR) is composed of several transmembrane protein subunits (there are 21 subunit isoforms known in the literature[22]) which form a chloride-ion (Cl^-) selective channel. Its function is initiated by the binding of GABA, considered as the principal inhibitory neurotransmitter of the central nervous system.

$GABA_A$/BzR agonists (anxiolytic, anticonvulsant, and sedative effects), inverse agonists (anxiogenic, stimulant, and convulsant effects), or antagonists (null efficacy) are recognized ligands which bond to GABA/BzR and enhance, diminish, or block the Cl^- channel, respectively. Several pharmacophore models have been proposed, that developed by Cook et al.[23,24] being relevant for the agonist and inverse agonist at the $GABA_A$/BzR with nonbenzodiazepine structures.

After this introduction, sections 2 and 3 of this paper describe the new contributions in the approximate similarity methodology and the QSAR models developed for GABA/BzR ligands, respectively. Results obtained with AS-based models are compared with those obtained using constitutional and fingerprints-based similarities. In addition, the new AS measurement is also compared with the 3D approach. Finally, conclusions of the new approaches developed and improvements achieved are exposed.

2. REFINING THE APPROXIMATE SIMILARITY (AS) MEASUREMENTS

2.1. Approximate Similarity Measurements. Given two molecules *A* and *B* represented by the G_A and G_B molecular graphs, respectively, the approximate similarity measurement, $AS_{A,B}$, reflects similarity and distance values based on graph isomorphism calculation and noncommon subgraphs, respectively, as follows:

$$AS_{A,B} = f(S_{A,B}, \Gamma_{A,B}, w_\Gamma) \quad (1)$$

where $S_{A,B}$ is the similarity obtained considering isomorphism ($I_{A,B}$) detection—approaches such as maximum common edges subgraph, maximum common subgraph (MCS), all maximum common subgraphs, fingerprints, and so on can be employed[25,26]—and different similarity metrics (Tanimoto, cosine, Raymond, etc.[26]); $\Gamma_{A,B}$ is the dissimilarity between the substructures which do not form $I_{A,B}$, and it is computed using topological invariants and any distance metric; w_Γ is a weighting factor which adjusts the distance contribution in the approximate similarity calculation; and finally, $f()$ is a mathematical function aimed at combining $S_{A,B}$, $\Gamma_{A,B}$, and w_Γ.

In a previous work,[15] different weights (w_Γ) were taken into account in classical similarity corrections as follows:

$$AS_{A,B} = S_{A,B} - w_\Gamma \bar{\Gamma}_{A,B} \quad (2)$$

where $S_{A,B}$ is the constitutional similarity value (considering the number of nodes and edges) and $\bar{\Gamma}_{A,B}$ is the $\Gamma_{A,B}$ value scaled by a normalization method with the aim of equaling both similarity and topological difference scales. The factor w_Γ—optimized for each chemical family (0.3 for steroids[15])—was constant in all of the comparisons of any pair of molecules of the data set.

However, a steroid-like AS approach could be not so adequate when the training and test compounds have less rigid and more dissimilar structures. New topological data treatments must be developed in order to obtain better AS descriptors. Thus, novel methods for calculating both nucleus similarity and fusing similarity and substituents dissimilarity were proposed. These methods are described below.

2.2. Similarity Based on Topological Invariants. Classical constitutional similarity does not take into account the characteristics of nodes and edges (e.g., color and weight), as can be observed in eq 3:

$$S^C_{A,B} = \frac{MCS_{A,B}}{\sqrt{A \times B}} \quad (3)$$

where $MCS_{A,B}$ is the number of nodes and edges of the maximum common subgraphs to the G_A and G_B molecular graphs and *A* and *B* represent the number of nodes and edges of the G_A and G_B graphs, respectively.

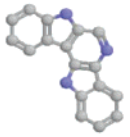
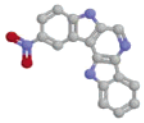
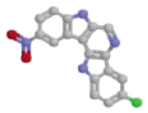
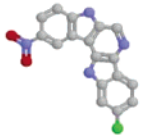
Thousands of topological invariants[27−29] have been proposed in the past 20 years in order to measure different graph characteristics and to correlate these characteristics with several chemical substance properties. The fact of including these topological descriptors in similarity calculation should refine the extracted chemical information and, in turn, improve the structure−activity relationships—the type and ratio of intramolecular bonds are employed instead of the number of nodes and edges.

For this purpose, once isomorphism has been detected by some of the approaches cited above, topological descriptors of G_A and G_B and common substructures which form the isomorphism are computed. Then, invariant-based similarity values can be obtained using the cosine index, as follows:

$$S^I_{A,B} = \frac{TD(MCS_{A,B})}{\sqrt{TD(A) \times TD(B)}} \quad (4)$$

where $TD(MCS_{A,B})$, $TD(A)$, and $TD(B)$ are the topological invariants of the maximum common substructures, of the *A* and *B* compounds, respectively.

Table 1. Comparisons between Molecules *A*, *B*, *C*, and *D*[a]

	A	B	C	D
	N = 20, E= 24	N= 23, E = 27	N = 24, E = 28	N = 24, E = 28
Descriptor	3725.93494	6250.9359	7427.8034	7458.6661
Activity	8.40	8.40	5.92	6.90
Molecules				
A		S^C = 0.9381 S^I = 0.7720 NIF = 0.0000	S^C = 0.9199 S^I = 0.7082 NIF = 0.0000	S^C = 0.9199 S^I = 0.7068 NIF = 0.0000
B	NIF = 2525.00		S^C = 0.9806 S^I = 0.9174 NIF = 0.0000	S^C = 0.9806 S^I = 0.9155 NIF = 0.0000
C	NIF = 3701.87	NIF = 1176.87		S^C = 0.9615 S^I = 0.8398 NIF = 1176.87
D	NIF = 3732.73	NIF = 1207.73	NIF = 1207.73	

[a] Numbers of nodes and edges (*N* and *E*) and HyperWiener descriptors (computed over the weighted matrices) of these molecules are shown. Constitutional and invariant-based similarities and NIF substructures, in addition to the biological activities, are also given.

Table 1 shows four molecules (*A*−*D*) which have a common substructure (the whole molecule *A*). Different information of the graph matching these molecules, in addition to their activities, is given.

When molecule *A* is compared with the remaining compounds, constitutional and invariant-based similarities are computed. $S^I_{A,B}$ shows more-different values of $S^I_{A,C}$ and $S^I_{A,D}$ than when $S^C_{A,B}$, $S^C_{A,C}$, and $S^C_{A,D}$ are employed. So, the invariant-based similarity tendency is in agreement with the variation of *A*, *B*, *C*, and *D* activities. A similar behavior is observed when molecule *B* is compared with *C* and *D*. A different substitution position of the chlorine produces a decrease of the S^I value regarding S^C, which agrees with the activity increase (5.92 for molecule *C* and 6.90 for the molecule *D*).

2.3. Measuring New Approximate Similarities. In a similar way to those works based on local fragment influence (local QSAR),[30,31] invariant-based similarity measurements should be corrected using nonisomorphic fragment (NIF) dissimilarities with the aim of achieving more realistic similarity values. Table 1 also shows the descriptor values for the NIF substructures extracted from each graph's pair matching between *A*−*D* molecules. Because molecule *A* is a substructure of molecules *B*, *C*, and *D*, $NIF_{A,B} = NIF_{A,C} = NIF_{A,D} = 0$ are obtained. A similar fact is observed when molecule *B* is compared with *C* and *D*. Furthermore, NIF matrices do not show a symmetric structure ($NIF_{i,j} \neq NIF_{j,i}$).

Thus, NIF information is used for correcting invariant-based similarities (*approximate similarity*) because external NIF substructures have key influence on activity values. In a previous work,[15] the contribution of nonisomorphic fragments in approximate similarity calculation was carried out by means of optimizing the factor w_Γ. This optimization—based on statistical parameters obtained in cross and external validation processes—provides the model with a constant weighting factor per data set.

Because the weighting factor was empirical, two shortcomings are involved, namely, w_Γ has not a priori chemical meaning and the optimization process can be time-consuming because of the procedural w_Γ search. In addition, the fact of considering a constant weighting factor does not involve an appropriate correction because the contribution of NIF substructures should depend on both the size and nature of the molecules to be compared and the value of the calculated isomorphism. Here, a new method for computing approximate similarities has been proposed and can be observed as follows:

$$AS_{A,B} = S^I_{A,B}\left[1 - \text{abs}\frac{TD(NIF_A) - TD(NIF_B)}{TD(A) + TD(B)}\right] \quad (5)$$

where $TD(NIF_A)$ and $TD(NIF_B)$ account for the NIF fragments of molecules *A* and *B*. In expression 5, the invariant-based cosine similarity index has been employed (other indexes have been tested and nonremarkable variation has been found). This similarity is corrected taking into consideration the size and nature of molecules *A* and *B* [values of TD(*A*) and TD(*B*)] and the dissimilarity of the nonisomorphic parts [values of $TD(NIF_A)$ and $TD(NIF_B)$].

In this way, similarity is modified using a function which informs about both isomorphic and nonisomorphic fragments. The greater the difference between $TD(NIF_A)$ and $TD(NIF_B)$ is, the greater the similarity correction is because the factor is closer to 0. As can be observed in expression 5, optimization of the contribution of nonisomorphic substructures in similarity is not empirically or manually modeled, so an automation of the approximate similarity computing is now possible.

3. RESULTS: BENZODIAZEPINE RECEPTOR LIGAND QSAR MODELS USING APPROXIMATE SIMILARITY VALUES

3.1. Data Sets. Figure 1 shows the nonbenzodiazepine $GABA_A$/BzR ligands and their experimental values (pIC_{50}) of inhibition of the [^{3}H]-diazepan-specific binding, which were extracted from the previous 3D-QSAR study of these ligands.[21] Experimental data of these compounds were obtained by the same pharmacological procedure, covering

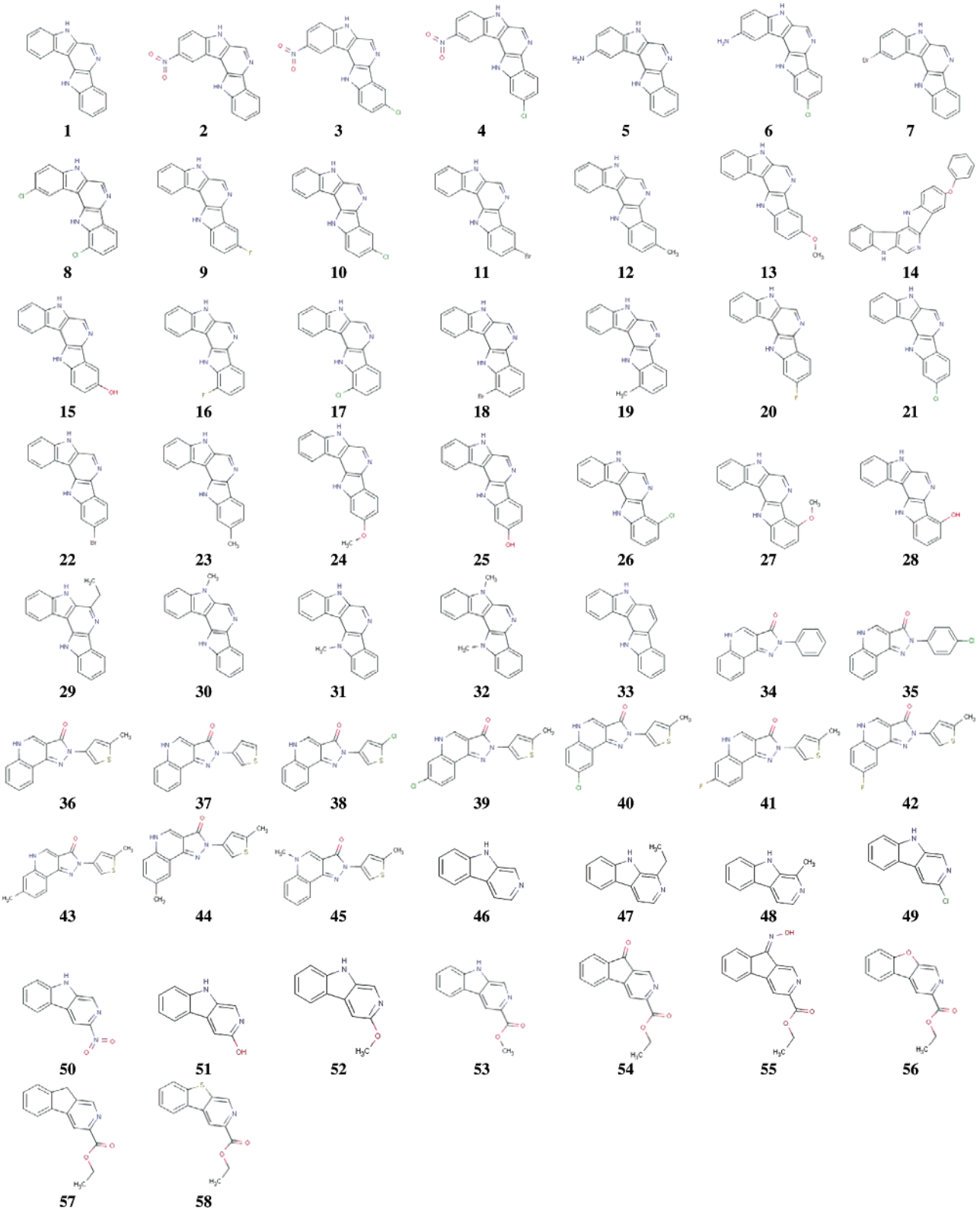


Figure 1. Nonbenzodiazepine ligands studied using approximate similarity: structures and lab activity values.

a wide biological activity range. In addition, an acceptable structural diversity was considered—there are three classes of compounds, namely, dihydroindolo-β-carboline structures (1−33), dihydro-pyrazolo-quinolinone compounds (34−45), and β-carboline ligands (46−58).

Ligand structures were divided into the training and test data sets in such a way that the use of 15% of the compounds for validation was ensured. The nine compounds employed by Verli et al.[21] for testing were also selected here, namely, compounds 9, 11, 16, 31, 32, 35, 44, 50, and 58. The training set was composed of the remaining ligands.

In addition, three approximate similarity matrices were considered:

- ^{2}AS: using expression 2 (several values for w_Γ have been tested, from 1.0 to 0.0, and $w_\Gamma = 0.35$ was the optimal weight).

•^{5}AS: using expression 5 and the HyperWiener[13,14,32,33] index considering weighted distance matrices where each (*i*,*j*) element represents the path length between the *i* and *j* atoms considering relative bond distances with regard to the C−C bond distance.

•^{5}AS*: using expression 5 and the HyperWiener index but considering in this case the classical distance matrix where all of the distances among atoms were considered as 1.0. The aim of this fact was to demonstrate the importance of the topological invariant selected.

Other topological descriptors based on weighted and nonweighted distance matrices have also been tested, but better results were not obtained because better correlations were not achieved.

On the other hand, similarity matrices built using constitutional similarity measurements and using fingerprints were employed with the aim of comparing the method here proposed with standard topological approaches. For this purpose, Chemaxon[34] software was employed for generating the similarity matrices on the basis of fingerprints. Constitutional similarity matrices were built using software developed by the authors.

3.2. Equations Training: Internal Validations and Study of Anomalies. Partial least-squares (PLS) regression was employed as a multivariate fitting technique because of its known characteristics, namely, data reduction is involved, variance of both the predictor matrix and biological activities is employed for model building, and predictors are considered as independent variables. The last characteristic enables the use of symmetric similarity matrices without algebraic restrictions.

The training stage was carried out by means of leave-one-out (LOO) validation, which consists of 49 individual cycles where the training and test sets were composed of 48 and 1 ligand, respectively, in each cycle. The fact of using each compound once for validation was ensured, and the final model was the average of the 49 individual cycles.

Figure 2 shows the lab versus predicted activity plots for the similarity approaches considered—^{2}AS, ^{5}AS and ^{5}AS* approximate similarity matrices and, on the other hand, constitutional and fingerprints-based similarity matrices—and their statistical parameters—the multiple determination coefficient (Q^2), the standard error in cross-validation (SECV), and the *slope* and bias (*intercept*); all of these referred to prediction—obtained in the training step.

As Figure 2 shows, the study of object−property outliers was also carried out in order to detect compounds which could introduce anomalies in the modeling stage (a noise component). The outlier removal strategy was to compute the *T* (student) parameter for all of the predicted values of the training set (residual divided by SECV) and to remove, if there were any, the compounds which show *T* values greater than 2.5. The cross-validation process was repeated until outliers were not found. The number of outliers was much lower than the limit allowed by the chemometric criterion (10−15% of the training set size).

The study of outliers must include the explanation of the anomalous behavior in order to justify the removal of these compounds. Structures 45, 47, and 53 of Figure 1 were detected as outliers for all of the similarity matrices (approximate, constitutional, and fingerprints approaches). As Table 2 shows, outliers 45 and 53 for the ^{2}AS and S^F

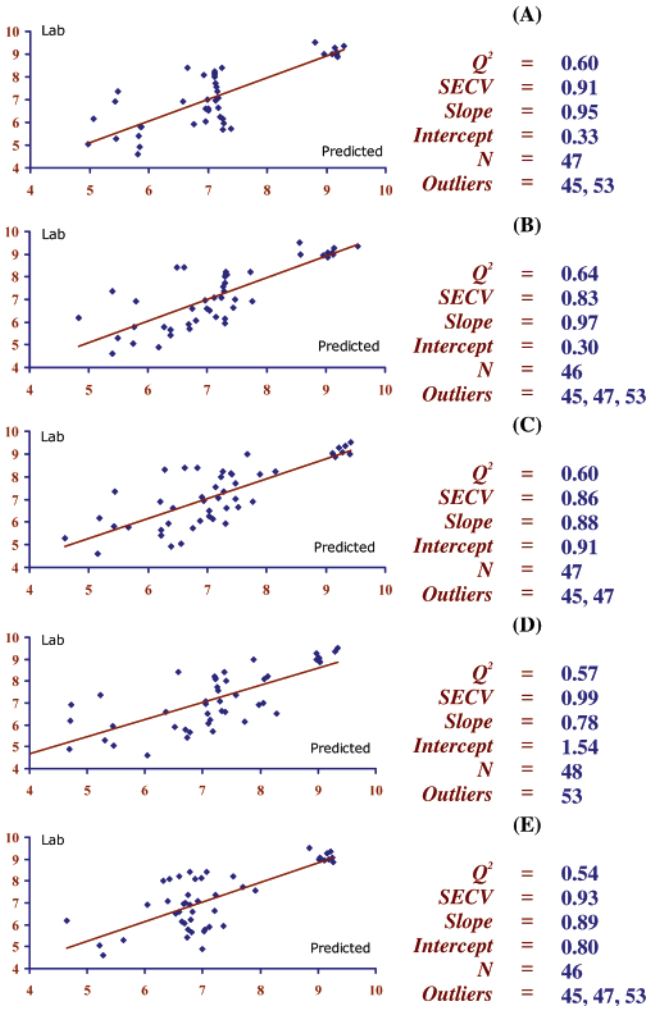


Figure 2. Lab vs predicted activity plots and statistical parameters of the training stages carried out using ^{2}AS (A), ^{5}AS (B), ^{5}AS* (C), constitutional similarity (D), and fingerprints-based similarity (E).

approaches were found; 45, 47, and 53 for ^{5}AS; 45 and 47 for ^{5}AS*; and only structure 45 for the S^C approach.

As can be observed in entry 45 of Figure 1, this ligand shows special characteristics compared with the rest of the compounds of its family (dihydro-pyrazolo-quinolinones): the methyl bonded to the amine nitrogen of the quinolinone substructure and its low lab activity. The latter can be explained with the pharmacophore model proposed by Cook et al.[23] This establishes that the amine nitrogen of the quinolinone interacts with an electron acceptor site in the receptor. Because the methyl does not favor this interaction, the activity is lower than those shown by the rest of the family. AS values did not model the high influence of this small structural difference in the activity prediction.

On the other hand, outliers 47 and 53 are the β-carboline ligands (46−58 nonbenzodiazepine subset) which show the lowest and greatest activities (3.60 and 8.30), respectively. Extrapolation could be the reason for the anomalous behavior of 47 and 53 because their activity values are very different from those shown by the rest of the β-carboline family. A model trained using more β-carboline ligands which shows extreme activities should predict correctly the 47 and 53 structures.

The ^{2}AS model only achieved screening levels: a distinction between low, medium, and high activities (three groups

Table 2. Lab and Predicted Activities (Using ^{2}AS, ^{5}AS, and ^{5}AS* Models) for the Training Set[a]

number	lab values	^{2}AS values	^{5}AS values	^{5}AS* values	S^C values	S^F values	3D values
1	8.40	7.24	6.61	6.83	7.37	7.07	7.93
2	8.40	6.65	6.49	6.62	6.58	6.62	7.92
3	5.92	6.76	6.68	6.34	6.51	7.17	5.83
4	6.90	6.57	7.75	7.77	7.97	6.81	7.08
5	7.36	7.16	7.28	7.27	7.57	7.15	7.51
6	7.01	7.00	7.47	7.47	8.06	6.79	7.13
7	8.22	7.11	7.31	7.26	7.21	7.52	8.09
8	6.51	7.01	7.04	7.03	7.08	6.70	6.73
10	5.67	7.24	6.38	6.21	6.78	6.93	5.88
12	6.65	7.18	7.44	7.52	7.33	7.19	7.46
13	6.05	6.97	6.81	6.88	7.10	6.69	6.56
14	5.80	5.88	6.27	5.42	3.65	6.75	5.58
15	6.94	7.14	6.97	6.94	7.06	6.67	7.06
17	7.10	7.16	7.11	6.91	7.06	6.59	6.68
18	7.55	7.14	7.27	7.13	7.26	7.91	7.80
19	7.08	7.17	7.24	7.18	7.30	6.90	7.90
20	8.15	7.11	7.32	7.40	7.22	7.00	7.73
21	8.00	7.12	7.30	7.22	7.40	6.52	7.53
22	7.72	7.12	7.31	7.47	7.25	7.68	7.67
23	8.10	7.11	7.31	7.42	7.22	6.87	7.56
24	8.10	6.93	7.34	7.89	8.07	6.43	7.41
25	8.22	7.12	7.73	8.15	8.13	6.60	7.78
26	6.15	7.25	7.30	7.09	7.72	6.80	6.64
27	6.60	7.01	6.75	6.42	6.36	6.60	5.75
28	6.24	7.21	7.15	7.02	7.12	6.78	7.76
29	6.60	6.96	6.99	7.32	7.39	6.81	6.42
30	5.94	7.27	7.29	7.31	5.44	7.33	6.21
33	5.72	7.39	6.70	6.76	7.17	7.04	7.30
34	9.51	8.82	8.55	9.42	9.35	8.83	8.92
36	9.35	9.30	9.55	9.32	9.29	9.22	9.08
37	8.99	9.10	9.13	9.40	8.97	9.19	8.67
38	8.99	8.96	8.57	7.67	7.88	9.03	8.76
39	8.96	9.18	8.95	9.13	9.02	9.12	9.09
40	9.06	9.17	9.11	9.28	9.01	9.05	8.96
41	9.05	9.17	9.02	9.11	9.01	9.25	8.85
42	9.29	9.15	9.13	9.23	8.97	9.16	9.11
43	8.87	9.19	9.03	9.15	9.04	9.26	9.19
45	6.52	outlier	outlier	outlier	8.29	outlier	6.46
46	5.79	5.86	5.77	5.66	6.70	7.06	5.62
47	3.60	5.85	outlier	outlier	4.38	outlier	3.59
48	4.91	5.84	6.18	6.38	4.69	6.97	4.60
49	7.35	5.48	5.40	5.44	5.22	6.31	6.59
51	5.40	5.84	6.37	6.22	6.73	6.75	6.41
52	6.91	5.44	5.80	6.21	4.72	6.10	5.86
53	8.30	outlier	outlier	6.28	outlier	outlier	6.93
54	4.59	5.82	5.40	5.14	6.04	5.30	5.77
55	5.30	5.45	5.48	4.59	5.31	5.64	4.67
56	5.04	4.98	5.75	6.55	5.46	5.23	6.53
57	6.17	5.08	4.83	5.17	4.71	4.68	6.51

[a] Predicted values using the constitutional and fingerprints-based models are also shown in addition to the 3D reference approach.

of compounds are clustered in the correlation plot, as shown in Figure 2A). The ^{5}AS and ^{5}AS* approaches led to better predictive correlations: $Q^2 = 0.64$ and SECV = 0.83 for ^{5}AS and $Q^2 = 0.60$ and SECV = 0.86 for ^{5}AS*. Approaches based on standard topological approaches showed a lower predictive ability: $Q^2 = 0.57$ and SECV = 0.99 for the constitutional similarity and $Q^2 = 0.54$ and SECV = 0.93 for the fingerprints-based similarity.

Table 2 shows the lab activities and the predicted values for the training set using the three AS approaches and the 3D QSAR model[21] from which experimental data were obtained and employed as a reference. In addition, predicted values using constitutional and fingerprints-based similarities are shown.

The effects of considering different families in a quality model were also studied. LOO processes were carried out separately for compounds 1−33 and 34−58. A different behavior was obtained by each one of the subfamilies considered: $Q^2 = 0.53$ and 0.81 for the 1−33 and 34−58 subsets, respectively (similar behavior is observed if referenced 3D data[21] are analyzed). So, the lower predictive ability achieved for the former subset could have hindered the excellent correlation of the latter. Because the target pursued was to develop a model with a wide applicability range, models covering a wide spectrum of nonbenzodiazepine ligands were attempted despite reducing the predictive power.

3.3. External Validation and Robustness Study. External validations of the above-constructed QSAR models were carried using the test set described in section 3.1 in order to evaluate the accuracy and robustness of the developed equations. Several statistical parameters—the correlation coefficient (r), the standard error in prediction (SEP), and the *slope* and bias (*intercept*)—were analyzed for the three approximate similarity approaches studied in this work: for the ^{2}AS model, $N = 9$, $r = 0.75$, SEP = 0.98, slope = 0.74, and intercept = 2.03; for the ^{5}AS model, $N = 9$, $r = 0.79$, SEP = 0.82, slope = 1.00, and intercept = 0.00; and for the ^{5}AS* model, $N = 9$, $r = 0.55$, SEP = 1.20, slope = 0.65, and intercept = 2.51.

The ^{5}AS model led to the best results according to all of the parameters. Thus, the robustness shown by the ^{5}AS values was maximal (SEP $\approx$ SECV) and confirms the fact of achieving a finer chemical representation using this approach. It is interesting to remark on the lower predictive ability obtained with the ^{5}AS* matrix, which also makes use of both invariant-based similarities (S^I) and variable correction factors, such as the ^{5}AS matrix. But topological descriptors were obtained from the nonweighted distance matrix, which depicts less chemical information and gives a nonrobust AS QSAR model.

Table 3 shows the lab activities and the predicted values using the three AS approaches and the 3D QSAR model[21] from which experimental data were obtained and employed as a reference. Results compare reasonably well and even improve the statistical parameters of the 3D QSAR ($r = 0.73$, SEP = 1.00, slope = 0.75, and intercept = 1.60).[21] It should be marked that the approximate model was built using topological measurements based on simple and fast methods.

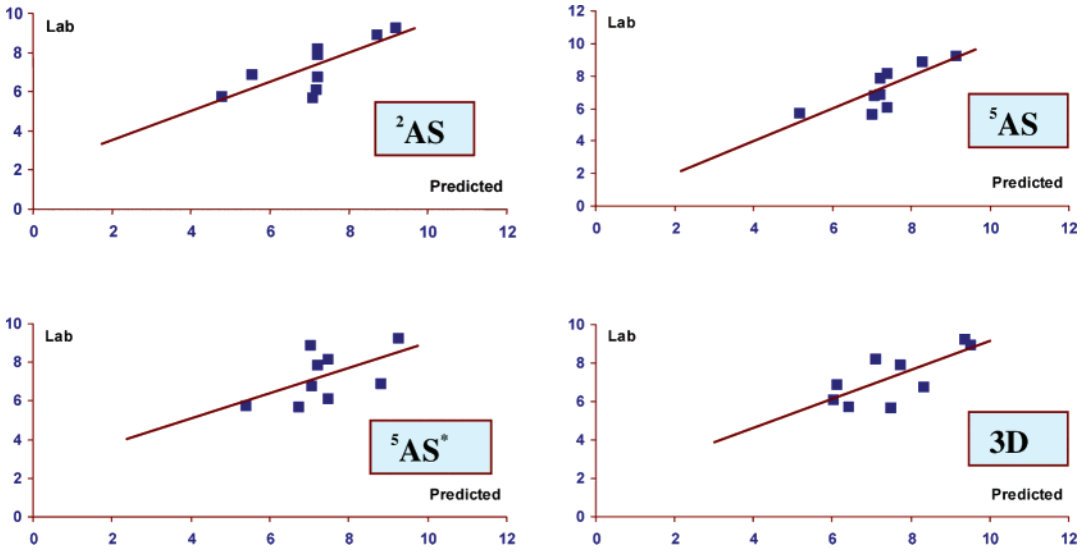
It is absolutely necessary to justify the fact of both studying and removing compounds which showed an anomalous behavior. For this purpose, the external predictions described above—predictions using equations without the outliers in the final equation modeling—were compared with those built using all of the training compounds—the detected outliers were employed in the training stage. For instance, the external validation carried out by the latter approach using the ^{5}AS model was as follows: $N = 9$, $r = 0.73$, SEP = 0.94, slope = 0.87, and intercept = 1.01.

Results obtained were not better than those considering the outliers' removal for training the ^{5}AS model. Similar tendencies were observed for the rest of the similarity approaches.

4. CONCLUSIONS

New approximate similarity methods have been proposed with the aim of building reliable topological QSAR models

Table 3. Lab and Predicted Activities (Using ^{2}AS and ^{5}AS and ^{5}AS* Models) for the Test Set[a]



number	lab values	^{2}AS values	^{5}AS values	^{5}AS* values	3D values
9	8.22	7.17	7.38	7.45	7.09
11	6.12	7.16	7.38	7.47	6.01
16	7.92	7.18	7.21	7.18	7.72
31	6.80	7.19	7.06	7.04	8.32
32	5.71	7.06	6.99	6.71	7.47
35	8.94	8.68	8.27	7.00	9.50
44	9.27	9.17	9.14	9.25	9.36
50	6.90	5.54	7.21	8.78	6.09
58	5.77	4.75	5.17	5.38	6.41

[a] Predicted values by the 3D reference approach are also shown.

for nonbenzodiazepine $GABA_A$/BzR ligands, which show high structural and activity variation. For this purpose, several developments have been studied, namely, (1) the modification of classical similarity indexes for considering topological invariants and (2) developments of new similarity correction processes which surpass the role of empirical weighting factors.

The fact of using topological descriptors provides more chemical information because isomorphism and noncommon substructures are depicted by pieces of chemical information which account for the number and nature of intramolecular bonds. Thus, topological descriptors were employed for similarity calculation and correction, and they were generated using weighted and nonweighted distance matrices. The former led to the best statistical parameters because the atoms and bond type are considered.

It is worthy to remark on the important characteristics achieved in the similarity correction process: the weighting factor depends directly on the size and nature of the molecular graphs compared, the amount of isomorphism detected, and the noncommon subgraphs. This allowed the achievement of lower or higher corrections in similarity as a function of the structures of the pair of molecules to be compared, and then, a better correlation with activities was obtained. In addition, this correction was implemented in a simple algorithm which automated the approximate similarity measurement.

The results achieved leave us to conclude that cheap and fast approximate similarity approaches can be developed to model and predict the pharmacological activity of a wide spectrum of nonbenzodiazepine $GABA_A$/BzR ligands. These QSAR tools were achieved despite the different behavior shown by the subfamilies of ligands considered.

The method here proposed compares reasonably well with a 3D-QSAR model and surpasses the results obtained with other standard topological approaches that consider characteristics of the functional groups of molecules (for instance, the use of fingerprints).

Future works should overtake the generation of more efficient similarities, weighting factor values, and the development of new isomorphism detection methods with the aim of obtaining better correlations. Studies considering topological descriptor combinations (considering electronegativities, charges, substituent effects, etc.), the influence of nonisomorphic fragments over the similarity and weight factor, and so forth are being conducted.

ACKNOWLEDGMENT

We thank the Comisión Interministerial de Ciencia y Technología (CiCyT) and FEDER for their financial support (Project: TIN2006-02071).

REFERENCES AND NOTES

(1) (a) Kubinyi, H. QSAR and 3D QSAR in Drug Design Part 1: Methodology. *Drug Discovery Today* **1997**, *2* (11), 457−467. (b) Kubinyi, H. QSAR and 3D QSAR in Drug Design Part 2: Applications and Problems. *Drug Discovery Today* **1997**, *2* (12), 538−546.
(2) *Structure−Property Correlations in Drug Research*; Van de Waterbeemd, H., Ed.; Academic Press: Austin, TX, 1996.
(3) Hansch, C. A. Quantitative Approach to Biochemical Structure−Activity Relationships. *Acc. Chem. Res.* **1969**, *2*, 232−239.

(4) *Handbook of Molecular Descriptors*; Todeschini, R., Consonni, V., Eds.; Wiley-VCH: Weinheim, Germany, 2000.
(5) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; John Wiley & Sons: New York, 2000.
(6) Greco, G.; Novellino, E.; Martin, Y. C. Approaches to Three-Dimensional Quantitative Structure−Activity Relationships. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 1997.
(7) *Nature-Inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks*; Leardi, R., Ed.; Elsevier: Amsterdam, 2003.
(8) Wold, S.; Sjostrom, M.; Eriksson, L. PLS−Regression: A Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.
(9) Golbraikh, A.; Tropsha, A. Be Aware of Q^2! *J. Mol. Graphics Modell*. **2002**, *20*, 269−276.
(10) Wold, S.; Eriksson, L. Statistical Validation of QSAR Results. In *Chemometrics Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, Germany, 1995; pp 309−318.
(11) Prabhakar, Y. S.; Gupta, M. K.; Roy, N.; Venkateswarlu, Y. A High Dimensional QSAR Study on the Aldose Reductase Inhibitory Activity of Some Flavones: Topological Descriptors in Modeling the Activity. *J. Chem. Inf. Model*. **2006**, *46*, 86−92.
(12) Almerico, A. M.; Tutone, M.; Lauria, A.; Diana, P.; Barraja, P.; Montalbano, A.; Cirrincione, G.; Dattolo, G. A Multivariate Analysis of HIV-1 Protease Inhibitors and Resistance Induced by Mutation. *J. Chem. Inf. Model*. **2006**, *46*, 168−179.
(13) Lučić, B.; Lukovits, I.; Nikolić, S.; Trinajstić, N. Distance-Related Indexes in the Quantitative Structure−Property Relationship Modeling. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 527−535.
(14) Ivanciuc, O.; Balaban, A. T. The Graph Description of Chemical Structures. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: Amsterdam, 1999; pp 59−167.
(15) Urbano Cuadrado, M.; Luque Ruiz, I.; Gómez-Nieto, M. A. A Steroids QSAR Approach Based on Approximate Similarity Measurements. *J. Chem. Inf. Model*. In press (proof available on Web).
(16) Cramer, R. D., III.; Paterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc*. **1988**, *110*, 5959−5967.
(17) Kubinyi, H. Comparative Molecular Field Analysis (CoMFA). In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; Wiley: New York, 1998.
(18) Golbraikh, A.; Tropsha, A. QSAR Modeling Using Chirality Descriptors Derived from Molecular Topology. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 144−154.
(19) Golbraikh, A.; Bonchev, D. Novel ZE-Isomerism Descriptors Derived from Molecular Topology and Their Application to QSAR Analysis. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 769−787.
(20) Yao, Y. Y.; Xy, L. Study on Structure−Activity Relationships of Organic Compounds: Three New Topological Indices and Their Applications. *J. Chem. Inf. Comput. Sci*. **1993**, *33*, 590−594.
(21) Verli, H.; Girão Albuquerque, M.; Bicca de Alencastro, R.; Barreiro, E. J. Local Intersection Volume: A New 3D Descriptor Applied to Develop a 3D-QSAR Pharmacophore Model for Benzodiazepine Receptor Ligands. *Eur. J. Med. Chem*. **2002**, *37*, 219−229.
(22) Olsen, R W.; DeLorey, T. M. GABA and Glycine. In *Basic Neurochemistry*; Siegel, G. J., Agranoff, B. W., Albers, R. W., Molinoff, P. B., Eds.; Raven Press: New York, 1999; pp 335−346.
(23) Huang, Q.; He, X.; Ma, C.; Liu, R.; Yu, S.; Dayer, C. A.; Wenger, G. R.; McKernan, R.; Cook, J. M. Pharmacophore/Receptor Models for GABAA/BzR Subtypes ($\alpha1\beta3\gamma2$, $\alpha5\beta3\gamma2$, and $\alpha6\beta3\gamma2$) via a Comprehensive Ligand-Mapping Approach. *J. Med. Chem.* **2000**, *43*, 71−95.
(24) Cox, E. D.; Diaz-Arauzo, H.; Huang, Q.; Reddy, M. S.; Ma, C.; Harris, B.; McKernan, R.; Skolnick, P.; Cook, J. M. Synthesis and Evaluation of Analogues of the Partial Agonist 6-(Propyloxy)-4-(methoxymethyl)-β-carboline-3-carboxylic Acid Ethyl Ester (6-PBC) and the Full Agonist 6-(Benzyloxy)-4-(methoxymethyl)-β-carboline-3-carboxylic Acid Ethyl Ester (Zk 93423) at Wild-Type and Recombinant GABAA Receptors. *J. Med. Chem.* **1998**, *41*, 2537−2552.
(25) 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*, 30−41.
(26) Willett, P. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(27) Randić, M. Topological Indices. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, U. K., 1998; pp 3018−3032.
(28) Balaban, A. T. Historical Developments of Topological Indices. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: Amsterdam, 1999; pp 21−57.
(29) Basak, S. C. Information Theoretic Indices of Neighborhood Complexity and Their Applications. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Delivers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: Amsterdam, 1999; pp 563−593.
(30) Sheridan, R. P.; Hunt, P.; Culberson, J. C. Molecular Transformations as a Way of Finding and Exploiting Consistent Local QSAR. *J. Chem. Inf. Model*. **2006**, *46*, 180−192.
(31) Lewis, R. A. A General Method for Exploiting QSAR Models in Lead Optimization. *J. Med. Chem.* **2005**, *48*, 1638−1648.
(32) Randić, M. Novel Molecular Descriptor for Structure−Property Studies. *Chem. Phys. Lett*. **1993**, *211*, 478−483.
(33) Klein, D. J.; Lukovtis, I.; Gutman, I. On the Definition of the Hyper-Wiener Index for Cycle-Containing Structures. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 50−52.
(34) Chemaxon Ltd. http://www.chemaxon.com/jchem (accessed May 2006).

CI600216H