Data Fusion of Similarity and Dissimilarity Measurements Using Wiener-Based Indices for the Prediction of the NPY Y5 Receptor Antagonist Capacity of Benzoxazinones

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

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

Received May 28, 2007

In this paper, we present the advantages of using data fusion of similarity and dissimilarity measurements for the development of quantitative structure−activity relationship models. Nonisomorphic fragments extracted in the matching process were considered to obtain dissimilarity values employed for correcting similarity measurements, thus, leading to finer chemical information. The purpose was to correlate similarity and dissimilarity matrices with pharmacological activities of drugs (the inhibitory capacity presented by 30 benzoxazinone derivatives for the NPY Y5 receptor). Wiener and hyper-Wiener descriptors computed over distance and weighted distance matrices were used for the calculation of dissimilarity values. A comparison with classical and fingerprints-based similarity was also carried out. The best approaches were achieved by means of dissimilarity and of fusion data spaces that take into account isomorphic and nonisomorphic information ($Q^2 = 0.88$, SECV = 0.18, slope = 1.06, and intercept = 0.09). The study of anomalous behavior presented by some compounds was also undertaken.

1. BACKGROUND

Quantitative structure−activity relationships (QSAR) represent an in silico methodology that establishes mathematical functions relating descriptions of biological receptor ligands with their pharmacological behavior.[1−5] In this way, estimation of ligand activities and parameters, which are only available after synthesizing drugs, can be carried out by structural descriptors obtained before drug development. So, synthesis processes are optimized with regard to environmental and economical factors.

One of the most widely used QSAR methods is based on obtaining the representation space through a similarity calculation among the data set elements.[6,7] The basic idea underlying similarity-based QSAR approaches was enunciated explicitly by Johnson and Maggiora,[8] who state that "molecules that are structurally similar likely will have similar properties". Hence, when the activity of a given molecule is unknown, we can predict it by taking into account similarity values between the molecule under study and the molecules of a data set whose activities are known. Currently, there is not agreement in defining quantitatively chemical similarity, and many methods have been proposed, each one showing different strengths and weaknesses. There is not a single method that is significantly better; however, obviously inappropriate methods exist.

Kolmogorov complexity[9] is a universal measure of similarity between two entities *A* and *B* (molecules in our context). The simplest measure is the length of the shortest program which takes *A* as input and produces *B* as output, that means the length of the shortest program that translates a representation into the other one. This quantity is called the conditional Kolmogorov complexity and is written K(B|A). But practical application of the Kolmogorov theory is very difficult. Recently, a new proposal of Kolmogorov complexity has been developed in order to measure differences between lengths of the files that store the compressed representations of two molecules by using text molecule representations (SMILE) and standard compression programs.[10]

One of the most employed similarity measurements is based on mapping fragments of a molecule into bits of a binary string (fingerprint).[11−13] It has been shown that fingerprints provide a nonintuitive encoding of molecular size, shape, and global similarity. Other common similarity measurements are based on the isomorphism calculation (e.g., maximum common substructure), and recently, similarity measurements which employ distances between topological descriptors computed over 2D molecular representations have also been proposed.[14−16] Through the use of these representations above, different similarity indices have been studied (e.g., Tanimoto, cosine, Raymond, etc.) to generate similarity values.[4]

In a general formulation, similarity measurements can be expressed as follows:[17]

$$S_{A,B} = f[g(\text{Structure}_A), g(\text{Structure}_B)] \qquad (1)$$

The application of this expression involves (a) molecular representations able to be handled by computers with the aim of matching each pair of molecules of a data set, (b) a function *g*(...) in charge of transforming structures into an amenable measurement if it is necessary (e.g., a molecular descriptor, fingerprints, etc.), and (c) the proposal of a

* Author to whom correspondence should be addressed. Phone: +34-957-21-2082. Fax: +34−957-21−8630. E-mail: ma1lurui@uco.es.
[†] University of Córdoba.
[‡] Institute of Chemical Research of Catalonia ICIQ.

10.1021/ci700180m CCC: $37.00
Published on Web 09/29/2007

function $f(...)$ to derive similarity measurements used later for correlating with experimental activity values in the following way: $h(\text{property}) = f[g(\text{Structure})]$.

Similarity descriptors often involve either 2D representations (topological descriptions) or 3D methodologies based on comparative molecular field analysis and comparative molecular similarity indices analysis approaches. In spite of obtaining good correlations between descriptors and activities for specific and reduced data sets, different descriptors account for different types of structural resemblance.[18,19] Therefore, some fusion approaches have been proposed by using several similarity measurements instead of a single one. The data fusion methodology involves the use of different similarity measurements and their experimental consensus scorings. In this way, similarity measurements are weighted in order to compute a similarity value showing richer chemical information by means of capturing relevant molecular descriptors related with the activity under study.

The search for better chemical similarity representations is justified by the current needs in the QSAR discipline: on one hand, top-level models should be useful in the prediction of activities of compounds not so structurally similar as those compounds employed in training and validation stages, and on the other hand, new QSAR methodologies should overcome problems related to the "*cliff*" phenomenon[20,21] (molecules that just have a slight structural variation and show quite different biological activities). The need for more complex descriptors should not involve drastic rises in resources and speed.

In this work, we present the use of similarity matrices built by means of correcting classical similarity values taking into account dissimilarity measurements over the noncommon parts of matched graphs. In this way, core substituents, which determine chemical activities many times, are considered in a direct way, thus, leading to better structure−activity correlations and detecting anomalous behavior. The methodology has been applied to the prediction of the NPY Y5 receptor antagonist capacity presented by a heterogeneous set of benzoxazinone derivatives.

After this introductory section, the translation (molecular description) and correlation (relationships with activity) functions are described in sections 2 and 3, respectively. Then, the experimental framework and a discussion about the predictive capability of different similarity approaches are given in sections 4 and 5. Finally, conclusions are presented.

2. SELECTION OF THE TRANSLATION FUNCTION

Several topological indices have been demonstrated to correlate with physical, chemical, or biological properties and activities of molecules.[1,3] The Wiener index (W) has been widely employed to develop quantitative structure−property relationship (QSPR) and QSAR models. This index was first introduced by Wiener[22] in 1947, but its graph-theoretical definition was pioneered by Hosoya[23] in 1971. The Wiener index is defined as follows:

$$W = \sum_{i<j}^{N} d_{ij} \qquad (2)$$

where d_{ij} is the shortest path between the nodes i and j of a molecular graph. The index counts distances between all the pairs of atoms in a molecule, so W is related to the molecular volume: the smaller W is, the larger the compactness of the molecule is.

This index was used to describe molecular branching and cyclicity and establish correlations with various physicochemical and thermodynamic parameters of chemical compounds, namely, boiling point, density, critical pressure, refractive indices, heats of isomerization and vaporization of various hydrocarbon species, and so forth.[24] The Wiener index has also found interesting applications in polymer chemistry, in studies of crystals, and in drug design and QSAR applications over a wide range of families of compounds (i.e., antiulcer agents, HIV-1 inhibitors, toxicity agents, and so on).

Many indices[1,4,7,25] have been devised from the Wiener index, and Randic's, Hosoya's, and Balaban's indices appear to be the most useful descriptors. In addition, some other indices based on adjacency and distance matrices have been also proposed (e.g., Zagreb group index, Platt's, Smolenskii, etc.). The hyper-Wiener (WW) index[26] is also derived from Wiener index and is defined as follows:

$$WW = \frac{1}{2}[\sum_{i<j}^{N} d_{ij}^{\ 2} + \sum_{i<j}^{N} d_{ij}] = \frac{1}{2}\sum_{i<j}^{N} d_{ij}^{\ 2} + \frac{1}{2}W \qquad (3)$$

Thus, the hyper-Wiener index is proportional to the Wiener index plus the sum of squared distances between pairs of vertices. The Wiener and hyper-Wiener indicies do not consider information about bonds and heteroatoms in their calculation. Recent modifications to these indices have been proposed by using weighted molecular graphs for their calculation. So, nonzero diagonal values and interatomic distances for edges are considered to obtain QSAR models showing a higher predictive ability.[27]

As commented above, Wiener-based indices have shown very good behavior for the description of many properties of compounds, and they have been widely used in the development of QSPR/QSAR models. Moreover, in previous studies, we have compared Wiener-based descriptors' efficiency against other more complex approaches; results obtained pointed out the better behavior of descriptors based on the distances and weighted distances.

We have developed chemoinformatic software for running similarity and dissimilarity calculations. This software includes a module for computing several Wiener-based descriptors and their more recent modifications (other kinds of indices are also managed by this software). Hence, the translation function employed in this work makes use of the computation of Wiener and hyper-Wiener descriptors computed over weighted isomorphic and nonisomorphic subgraphs.

3. SELECTION OF THE CORRELATION FUNCTION

A large number of mathematical functions correlating similarity and activity values can be established. The correct choice of the correlation function depends on, among other factors, the extent of the activity variation with regard to structural characteristics.

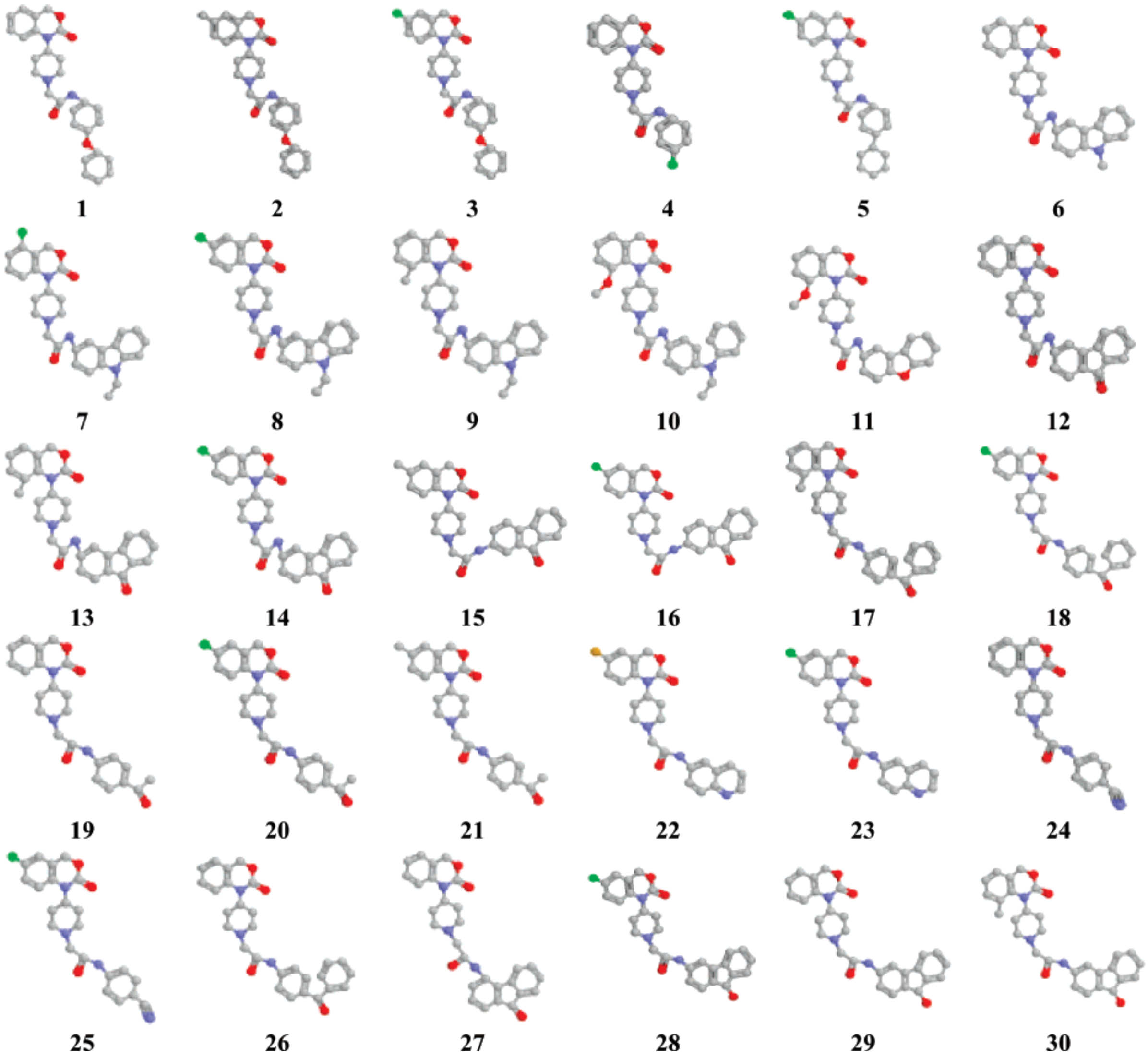

Figure 1. Structures of the 30 benzoxazinone derivatives employed in this work.

Linearity of the relationship between similarity and activity is required to apply traditional multivariate regression methods, namely, multivariate linear regression, principal component regression, and partial least-squares regression (PLSR).[7,28−30] When the relationship does not show a linear modeling, nonparametric methods can be attempted, namely, artificial neural networks composed by a set of layers of neurons that deals with input information (similarity space) through bioinspired modeling and generates output information (activity space). Hence, feature selection methods can be employed to select specific variables giving information about local fragments which determine molecular activity.

PLSR was employed in this work as a regression technique. Matlab[31] code implemented by the authors was used to develop leave-one-out (LOO) validation processes, and statistical parameters referring to prediction ability (not fitting) were computed, namely, the ecoefficient of determination (Q^2), standard error in cross-validation (SECV), and the slope and intercept of the experimental versus predicted plot. External validations were also carried out by using compounds not employed in the model training with the aim of evaluating the predictive capability of models.

In spite of not employing qualitative approaches in this research, pattern recognition methods could be useful for predicting if compounds show high or low activity. The analysis of this kind of models is simpler than the study of quantitative QSAR equations.

4. EXPERIMENTAL FRAME

4.1. Data Set. Figure 1 shows the structures of the 30 benzoxazinone derivates selected as the chemical space to be modeled. The NPY Y5 receptor antagonist capacity was the activity studied for this data set. Pharmacological studies point out that NPY Y5 antagonists are potential antiobese agents.[32] It is widely accepted that obesity influences many kinds of diseases and disorders, namely, respiratory, musculoskeletal, gastrointestinal, cardiovascular, and so forth. So, the development of effective and safe antiobesity drugs is of great interest to pharmacological chemists.

The inhibitory activity (IC_{50}) of the data set has been compiled from the literature.[32] Table 1 shows the data set SMILE structures and their $pIC_{50} = -\log(1/IC_{50})$ values. The statistical information of the data set is as follows: $N = 30$, mean $= 1.64$, min $= 0.88$, max $= 2.88$, standard deviation $= 0.51$.

4.2. Data Modeling and Calculation. Graph structures and keyed fingerprints were generated by using the *Marv-*

Table 1. SMILE Structures of the Benzoxazinone Derivatives and Their Values of Inhibitory Capacity as NPY Y5 Receptor Antagonist

	molecules	pIC_{50}
1	O=C(CN1CCC(CC1)N2C(=O)OCC3=CC=CC=C23)NC4=CC=C(OC5=CC=CC=C5)C=C4	1.30
2	CC1=CC=C2N(C3CCN(CC3)CC(=O)NC4=CC=C(OC5=CC=CC=C5)C=C4)C(=O)OCC2=C1	2.02
3	ClC1=CC=C2N(C3CCN(CC3)CC(=O)NC4=CC=C(OC5=CC=CC=C5)C=C4)C(=O)OCC2=C1	1.78
4	ClC1=CC=C(NC(=O)CN2CCC(CC2)N3C(=O)OCC4=CC=CC=C34)C=C1	2.48
5	ClC1=CC=C2N(C3CCN(CC3)CC(=O)NC4=CC=C(C=C4)C5CCCCC5)C(=O)OCC2=C1	2.05
6	CN1C2=CC=CC=C2C3=C1C=CC(NC(=O)CN4CCC(CC4)N5C(=O)OCC6=C5C=CC=C6)=C3	0.98
7	CCN1C2=CC=CC=C2C3=C1C=CC(NC(=O)CN4CCC(CC4)N5C(=O)OCC6=C5C=CC=C6Cl)=C3	1.74
8	CCN1C2=CC=CC=C2C3=C1C=CC(NC(=O)CN4CCC(CC4)N5C(=O)OCC6=C5C=CC(Cl)=C6)=C3	2.00
9	CCN1C2=CC=CC=C2C3=C1C=CC(NC(=O)CN4CCC(CC4)N5C(=O)OCC6=C5C(C)=CC=C6)=C3	1.70
10	CCN(C1=CC=CC=C1)C2=CC=C(NC(=O)CN3CCC(CC3)N4C(=O)OCC5=C4C(OC)=CC=C5)C=C2	2.88
11	COC1=CC=CC2=C1N(C3CCN(CC3)CC(=O)NC4=CC5=C(OC6=CC=CC=C56)C=C4)C(=O)OC2	1.75
12	O=C(CN1CCC(CC1)N2C(=O)OCC3=C2C=CC=C3)NC4=CC5=C(C=C4)C(=O)C6=CC=CC=C56	1.37
13	CC1=CC=CC2=C1N(C3CCN(CC3)CC(=O)NC4=CC5=C(C=C4)C(=O)C6=CC=CC=C56)C(=O)OC2	1.70
14	ClC1=CC2=C(C=C1)N(C3CCN(CC3)CC(=O)NC4=CC5=C(C=C4)C(=O)C6=CC=CC=C56)C(=O)OC2	1.40
15	CC1=CC2=C(C=C1)N(C3CCN(CC3)CC(=O)NC4=CC5=C(C=C4)C6=CC=CC=C6C5=O)C(=O)OC2	1.94
16	ClC1=CC2=C(C=C1)N(C3CCN(CC3)CC(=O)NC4=CC5=C(C=C4)C6=CC=CC=C6C5=O)C(=O)OC2	1.84
17	CC1=CC=CC2=C1N(C3CCN(CC3)CC(=O)NC4=CC=C(C=C4)C(=O)C5=CC=CC=C5)C(=O)OC2	1.60
18	ClC1=CC2=C(C=C1)N(C3CCN(CC3)CC(=O)NC4=CC=C(C=C4)C(=O)C5=CC=CC=C5)C(=O)OC2	1.05
19	CC(=)C1=CC=C(NC(=O)CN2CCC(CC2)N3C(=O)OCC4=C3C=CC=C4)C=C1	1.24
20	CC(=)C1=CC=C(NC(=O)CN2CCC(CC2)N3C(=O)OCC4=C3C=CC(Cl)=C4)C=C1	1.32
21	CC(=)C1=CC=C(NC(=O)CN2CCC(CC2)N3C(=O)OCC4=C3C=CC(C)=C4)C=C1	1.41
22	FC1=CC2=C(C=C1)N(C3CCN(CC3)CC(=O)NC4=CC5=CC=CN=C5C=C4)C(=O)OC2	0.89
23	ClC1=CC2=C(C=C1)N(C3CCN(CC3)CC(=O)NC4=CC5=CC=CN=C5C=C4)C(=O)OC2	2.17
24	O=C(CN1CCC(CC1)N2C(=O)OCC3=C2C=CC=C3)NC4=CCC(C=C4)C#N	2.12
25	ClC1=CC2=C(C=C1)N(C3CCN(CC3)CC(=O)NC4=CCC(C=C4)C#N)C(=O)OC2	2.30
26	O=C(CN1CCC(CC1)N2C(=O)OCC3=C2C=CC=C3)NC4=CC=C(C=C4)C(=O)C5=CC=CC=C5	0.88
27	O=C(CN1CCC(CC1)N2C(=O)OCC3=C2C=CC=C3)NC4=CC=CC5=C4C6=CC=CC=C6C5=O	2.14
28	OC1C2=CC=CC=C2C3=C1C=CC(NC(=)CN4CCC(CC4)N5C(=)OCC6=C5C=CC(Cl)=C6)=C3	0.90
29	OC1C2=CC=CC=C2C3=C1C=CC(NC(=O)CN4CCC(CC4)N5C(=O)OCC6=C5C=CC=C6)=C3	0.94
30	CC1=CC=CC2=C1N(C3CCN(CC3)CC(=O)NC4=CC5=C(C=C4)C(O)C6=CC=CC=C56)C(=O)OC2	1.48

inSketch software and *generfp* (default options) of the Jchem package, respectively.[33] Structural isomorphism was extracted by using an algorithm developed by the authors,[34] considering the maximum common substructure (MCS) as the isomorphic fragment (isomorphic and nonisomorphic substructures were retained for similarity calculations). When fingerprints were employed, isomorphic fragments were represented by those bits set to 1 and common to the two compared fingerprints (each 1 or 0 bit accounts for the presence or nonpresence of a specific subgraph, respectively).

Regarding similarity indices, Tanimoto and cosine formulas were those studied in this work. These indices are shown in expressions 4 and 5, where L_A and L_B are the sizes of the compared graphs (G_A and G_B), while L_C accounts for the isomorphic fragment.

$$\text{Tanimoto index} = S_{A,B} = \frac{L_C}{L_A + L_B - L_C} \tag{4}$$

$$\text{cosine index} = S_{A,B} = \frac{L_C}{\sqrt{L_A L_B}} \tag{5}$$

For dissimilarity calculations, Wiener and hyper-Wiener descriptors were computed over complete graphs and nonisomorphic fragments by using both distance and weighted-distance graph matrices. On weighted distance matrices, distances equal to 1 between the connected nodes *i* and *j* were replaced by bond lengths between the connected atoms *i* and *j* relative to the distance to the bond $C-C$. Software developed by the authors was used for this calculation.

Hence, similarity, dissimilarity, and fused matrices were built and employed as multivariate spaces. These 30 × 30 matrices can be considered as sets of 30 objects (rows) characterized by 30 variables (columns) which inform about the similarity or dissimilarity between the compound (object) and a reference compound. As validation strategies, internal (LOO) and external tests were carried out. Regarding the latter, several test subsets were randomly selected.

Table 2. Statistical Results and the Number of Property-Descriptor Outliers Obtained in LOO Processes for Constitutional and Fingerprint-Based Similarity Values Computed by Using Tanimoto and Cosine Indices

	index	Q^2	slope	bias	SECV	outliers
constitutional	Tanimoto	0.83	0.89	0.16	0.18	7
	cosine	0.73	0.87	0.20	0.22	6
fingerprint-based	Tanimoto	0.80	0.96	0.09	0.23	5
	cosine	0.75	0.79	0.33	0.28	5

5. EXPERIMENTAL RESULTS

5.1. Classical Similarity Analysis. First, constitutional (number of common nodes and edges) and fingerprint-based (number of common bits set to 1 representing the same structural characteristic) similarity matrices were built in order to analyze the results obtained by means of considering only isomorphic information. Table 2 shows the statistical results and the number of property-descriptor outliers obtained in LOO processes. The outliers study was carried out by setting a cutoff value for the *T* parameter, which is computed as the ratio between the deviation of the predicted activity obtained for an object (molecule) and the error obtained in prediction (in our case, SECV). This is a way to detect those samples that show an anomalous behavior with regard to the rest of the data set. A $T_{\text{cut-off}}$ value set to 2.5 is often used in multivariate models.[35] As can be observed in Table 2, in spite of obtaining good correlations once outliers had been removed, the number of outliers was

Table 3. Statistical Results and the Number of Property-Descriptor Outliers Obtained in LOO Processes for Dissimilarity Values Computed by Wiener and Hyper-Wiener Indices over Distance and Weighted Distance Matrices

matrix	index	Q^2	slope	bias	SECV	outliers
distance	Wiener	0.51	0.84	0.25	0.36	1
	hyper-Wiener	0.77	1.02	0.05	0.24	3
weighted distance	Wiener	0.50	0.84	0.26	0.36	1
	hyper-Wiener	0.77	1.02	0.06	0.24	3

excessive (it cannot be greater than 10% of the data set size). This behavior of classical similarity methods could be due to not considering directly the nonisomorphic data in modeling.

5.2. Nonisomorphic Fragments Behavior. Several works have been proposed by taking into account differences between nonisomorphic fragments extracted from matching processes with the aim of modeling different data sets.[36−38] A dissimilarity measure can be obtained by means of using an appropriate descriptor over molecular graphs that correspond to nonisomorphic fragments (fully connected graphs are not required).

If the matching algorithm proposed by the authors[34] is applied to the molecules of a data set given, an isomorphic fragment ($I_{A,B}$) and two non-necessarily connected nonisomorphic substructures (NIF_A and NIF_B) are obtained for each pair of molecules A and B. A dissimilarity or distance value can be obtained as follows:

$$\bar{d}_{A,B} = \frac{\sqrt{{\mathrm{TD}_{\mathrm{NIF}_A}}^2 + {\mathrm{TD}_{\mathrm{NIF}_B}}^2}}{\mathrm{TD}_A \times \mathrm{TD}_B} \tag{4}$$

where TD_A and TD_B represent the descriptors computed over the molecules A and B, respectively, and $\mathrm{TD}_{\mathrm{NIF}_A}$ and $\mathrm{TD}_{\mathrm{NIF}_B}$ are the descriptor values for the nonisomorphic structures of A and B. Thus, $\bar{d}_{A,B}$ is a dissimilarity value which accounts for the nonisomorphic fragments of the matched molecules. Table 3 shows results obtained by using dissimilarity matrixes (Wiener and hyper-Wiener were employed).

As observed in Table 3, there is no difference between using distance or weighted-distance matrices. When the hyper-Wiener index was considered, excellent correlations were obtained ($Q^2 > 0.75$, slope $= 1.0$, bias $= 0.0$) and the number of outliers was reduced to three compounds (**1**, **6**, and **22**). Nevertheless, if the Wiener invariant is employed, statistical parameters are worse in spite of obtaining only one outlier (molecule **22**).

From the results commented upon above, we can conclude that the consideration of squared terms by means of the hyper-Wiener definition involves a refinement of the nonisomorphic structure-based influence on the distance measure.

5.3. Combining Isomorphic and Nonisomorphic Information. Several trends can be derived from the study of the models above: (a) better results were obtained by using the Tanimoto index and the hyper-Wiener invariant, and (b) there are not important differences between the results when using distances and weighted-distances matrices. So, these parameters were taken into account for the subsequent developments.

Despite having high values of Q^2, classical similarity-based models were not appropriate due to the excessive number of outliers (problems related to robustness): molecules **4**, **8**, **10**, **11**, **22**, **23**, and **27** when the Tanimoto index and the constitutional similarity were involved and molecules **4**, **7**, **11**, **22** and **23** when the Tanimoto index and fingerprints were employed. These compounds that show anomalous behavior were also detected when other indices were considered. Molecule **22** was also detected in nonisomorphic-based QSAR developments.

We can classify outliers into two types, namely, (a) those which do not belong to the chemical space to be modeled (structural or activity space), thus, requiring new objects to cover the non-well-defined chemical regions, and (b) outliers characterized by the *cliff* phenomenon, which is derived from great activity variations resulting from small structural changes. So, solutions to the *cliff* outliers are difficult to solve.

The detection of molecules **22** and **23** as outliers is due to the *cliff* problem since these two molecules are extremely similar and show quite different activities, as can be observed in Table 1 and Figure 1. Furthermore, other outliers (**4** and **10**) show problems related to the incorrect definition of the activity space that is predicted, as shown in the histogram of Figure 2. We can observe that there are some intervals only defined by one object (one compound), thus, leading to poor predictions for these molecules.

As stated in the *background* section, several solutions have been proposed in order to build appropriate predictive spaces by similarity-based QSAR methods. They employed those descriptors showing high correlations with the activity under study and different similarity measures weighted by optimal *consensus* factors. In previous works,[15,36,37] we have shown the usefulness of the approximate similarity (AS) concept to develop QSAR models. Approximate similarity is based on correcting constitutional similarities by means of distance or dissimilarity values corresponding to the nonisomorphic fragments extracted from matching processes. Dissimilarity values are generated by the calculation of a topological descriptor over molecular graphs that represent the nonisomorphic substructures. The use of fused information (isomorphic and nonisomorphic data) could be useful for enlarging differences between similar compounds showing different properties, but a strict study of the consensus merging function must be carried out in order to avoid enlarging differences between similar compounds showing similar properties.

Thus, AS is defined as follows:

$$\mathrm{AS}_{A,B} = f(S_{A,B},\Gamma_{A,B},w_\Gamma) \tag{5}$$

where $S_{A,B}$ is a classical similarity measurement (constitutional, fingerprint-based, or descriptor-based), $\Gamma_{A,B}$ is the distance or dissimilarity measurement obtained through use of the nonisomorphic fragments, and w_Γ is a weighting factor which adjusts the contribution of the nonisomorphic fragments to the similarity measurement.

The choice of an appropriate function f and optimization of the distance contribution to the similarity correction must be carried out for each data set depending on the predictive ability. For benzoxazinone derivatives, the AS expression was defined as follows:

$$\mathrm{AS}_{A,B} = S_{A,B} - \left[\bar{d}_{A,B} - \mathrm{abs}\left(\frac{\mathrm{TD}_{\mathrm{MCS}}}{\mathrm{TD}_A} - \frac{\mathrm{TD}_{\mathrm{MCS}}}{\mathrm{TD}_B}\right)\right] \tag{6}$$

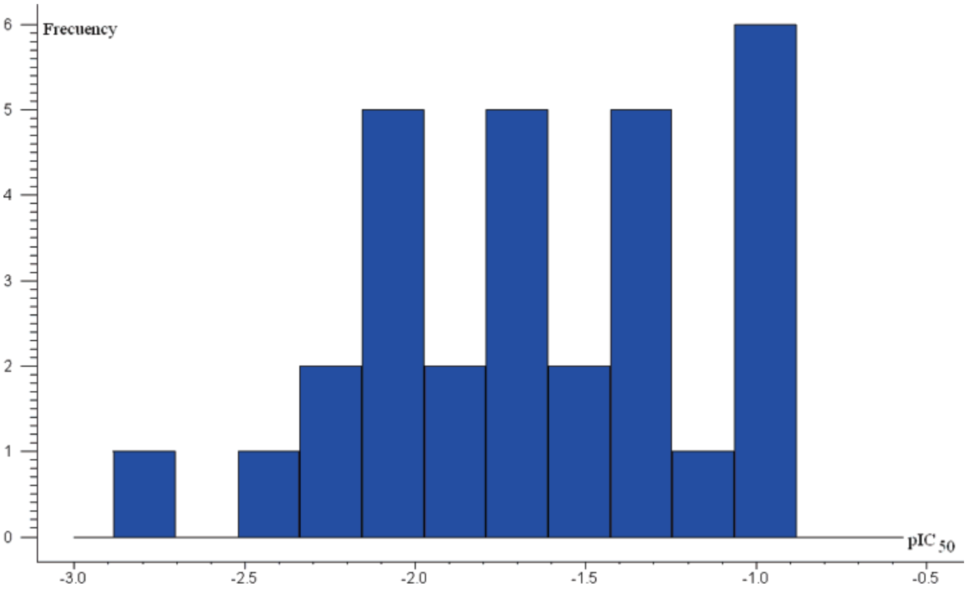

Figure 2. Frequency histogram for the pIC_{50} values of the 30 benzoxazinone derivatives.

Table 4. Results Obtained for Five QSSR Models Built by Using Approximate Similarity Values and Tested by five External Sets Generated by a Systematic Generator[a]

test set	fitting r^2	test r^2	test set composition and predictions
1	1.00	0.87	1(1.30, 1.74), 6(0.98, 1.11), 11(1.75, 1.92), 17(1.60, 1.57), 23(2.17, 1.98), 28(0.90, 0.62)
2	1.00	0.35	2(2.02, 1.34), 7(1.74, 1.82), 12(1.36, 1.48), 18 (1.05, 1.56), 24(2.12, 2.10), 29(0.94, 1.62)
3	1.00	0.83	3(1.78, 1.20), 8(2.00, 1.91), 13(1.70, 1.35), 19(1.24, 1.27), 25(2.30, 2.14), 30(1.48, 0.89)
4	0.99	0.98	4(2.48, 2.30), 9(1.70, 1.72), 14(1.40, 1.56), 20(1.32, 1.31), 26(0.88, 0.85)
5	0.99	0.90	5(2.05, 2.48), 10(2.88, 2.81), 16(1.84, 1.57), 21(1.41, 1.43), 27(2.14, 2.08)

[a] Fitting was carried out by the remaining benzoxazinone derivatives. (The first number is the compound number shown in Figure 1, and the numbers in brackets are the lab and predicted values for the test compounds.)

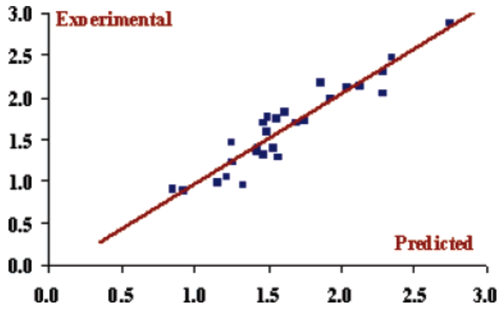

Figure 3. Predicted vs experimental pIC_{50} values plot for the approximate similarity values.

where TD_{MCS} is the value of the hyper-Wiener index corresponding to the isomorphic fragment extracted from the matching of A and B molecules and $S_{A,B}$ is the constitutional similarity using the Tanimoto index, and the remaining terms have been defined previously. Thus, the correcting term considered descriptions of nonisomorphic subgraphs and differences between A and B graphs with regard to the MCS value.

Figure 3 shows the representation of experimental versus predicted pIC_{50} values obtained with expression 6. Statistical characterization of the prediction capacity was as follows: $Q^2 = 0.88$, SECV = 0.18, slope = 1.06, and intercept = 0.09.

Therefore, the accuracy and precision of the predictions carried out were significantly improved, leading to useful uncertainty reductions for computer-aided drug development. Three outliers were detected, namely, compounds **2**, **15**, and **22**. As can be observed, outlier **22** is common to all of the models (the *cliffs* problem).

External validations were carried out in order to evaluate the predictive capability of models built by means of the similarity-corrected methods (AS) for compounds which were employed in training. Independent sets of molecules were considered by selecting randomly several training and test subsets (a systematic generator was employed for selecting six different test subsets). Experimental and predicted activity values for the approximate similarity approach are shown in Table 4. As can be observed, the size of the test subsets (five or six compounds) ensured an appropriate number of test molecules. The composition of the different test sets and their r^2 parameters obtained in the fitting and test stages are given. The systematic generator, in addition to the fact of providing random subsets, allowed for all of the compounds to be employed for testing at least once.

Good results were obtained for all the models with the exception of the second one, whose test r^2 value is quite poor. The four remaining models showed test r^2 values higher than 0.80, thus, pointing to the high robustness achieved by the similarity-corrected modeling proposed in this paper.

6. CONCLUSIONS

In this work, we have studied several graph-based methods to develop QSAR models with the aim of predicting the inhibitory capacity presented by 30 benzoxazinone derivatives for the NPY Y5 receptor. Isomorphic and nonisomorphic information were first employed, and better correlations were obtained by means of using dissimilarity data accounting for the differences between molecules. But several compounds were detected as samples showing anomalous behavior owing to the *cliff* phenomenon or to problems related to the chemical space definition. The lowest number of outliers was obtained for nonisomorphic matrices: three compounds showed anomalous deviations, this number being lower than the number of outliers accepted by the chemometric community as a cutoff for the development of predictive models (15% of the data set size).

Data fusion of the descriptors computed over isomorphic and nonisomorphic subgraphs was also attempted, and good results were obtained. So, we can conclude that QSAR development, most of the time, requires the use of different kinds of information in order to build reliable tools. The merging function and the contributions of the two kinds of data employed must be optimized for each chemical family to be modeled.

It is interesting to remark on the employment of fast and cheap tools to develop the QSAR models here presented since only 2D computations are involved and geometry optimization and alignment are not required.

ACKNOWLEDGMENT

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

REFERENCES AND NOTES

(1) Diudea, M. V. *QSPR/QSAR Studies by Molecular Descriptors*; Nova Science Publishers: Huntington, NY, 2001.

(2) Guha, R.; Jurs, P. C. Determining the Validity of a QSAR Model − A Classification Approach. *J. Chem. Inf. Model.* **2005**, *45*, 65−73.

(3) Hansch, C.; Leo, A. *Exploring QSAR: Fundamental and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.

(4) *Topological Indices and Related Descriptors in QSAR and QSPR*. Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: Amsterdam, 2000.

(5) van de Waterbeemd, H.; Gifford, E. ADMET in Silico Modelling: towards Prediction Paradise? *Nat. Rev. Drug Discovery* **2003**, *2*, 192−204.

(6) Nikolova, N.; Jaworska, J. Approaches to Measure Chemical Similarity − A Review. *QSAR Comb. Sci.* **2004**, *22*, 1006−1026.

(7) *Chemoinformatics: A Textbook*; Gasteiger, J., Engel, T., Eds.; Wiley-VCH: Weinheim, Germany, 2003.

(8) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; John Wiley: New York, 1990.

(9) Li, M.; Vitanyi, P. *An introduction to Kolmogorov complexity and its applications*; Springer-Verlag: New York, 1997.

(10) Melville, J. L.; Riley, J. F.; Hirst, J. D. Similarity by Compression. *J. Chem. Inf. Model.* **2007**, *47*, 25−33.

(11) *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, G., Eds.; JAI Press: London, 1998; Vol. 2.

(12) Flower, D. R. On the Properties of Bit String-based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.

(13) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569−574.

(14) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.

(15) 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.

(16) Nikolova, N.; Jaworska, J. Approaches to Measure Chemical Similarity − a review. *QSAR Comb. Sci.* **2003**, *22*, 1006−1026.

(17) Bender, A.; Jenkins, J. L.; Li, O.; Adams, S. E.; Cannon, E. O.; Glen, R. C. Molecular Similarity: Advances in Methods, Applications and Validations in Virtual Screening and QSAR. In *Annual Reports in Computational Chemistry*; Elsevier B.V.: New York, 2006; Vol. 2.

(18) Holliday, J. D.; Jelfs, S. P.; Willett, P.; Gedeck, P. Calculation of Intersubstituent Similarity using R-Group Descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 406−411.

(19) Holliday, J. D.; Salim, N.; Whittle, M.; Willett, P. Analysis and Display of the Size Dependence of Chemical Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 819−828.

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

(21) 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.

(22) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.

(23) Hosoya, H. Topological Index: A Newly Proposed Quantity Characterizing the Topological Nature of Structured Isomers of Saturated Hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332−2339.

(24) Gutman, I.; Klavzar, S.; Mohar, B. Fifty Years of the Wiener Number. *MATCH* **1997**, *35*.

(25) *Handbook of Molecular Descriptors*; Todeschini, R., Consonni, V., Eds.; Wiley-VCH: Weinheim, Germany, 2000.

(26) Klein, D. J.; Lukovits, I.; Gutman, I. On the Definition of the Hyper-Wiener Index for Cycle-Containing Structures. *MATCH* **1995**, *35*, 50−52.

(27) Ivanciuc, O.; Ivanciuc, T.; Balaban, A. T. Vertex- and Edge-Weighted Molecular Graphs and Derived Structural Descriptors. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon & Breach Science Publishers: The Netherlands, 1999; pp 169−220.

(28) Rouvray, D. H. In *Mathematics and Computational Concepts in Chemistry, Proceedings of International Symposium on Applications of Mathematical Concepts to Chemistry*; Trinajstic, N., Ed.; Ellis Horwood: Chichester, U.K., 1986.

(29) Wold, S.; Sjostrom, M.; Eriksson, L. PLS-Regression: A Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.

(30) *MultiVariate Data Analysiss in Practice*; Esbensen, K. H., Ed.; Camo Process AS: Oslo, 2002.

(31) The MathWorks Inc. http://www.mathworks.com (accessed March 2007).

(32) Deswal, S.; Roy, N. Quantitative structure activity relationship of benzoxazinone derivatives as neuropeptide Y Y5 receptor antagonists. *Eur. J. Med. Chem.* **2006**, *41*, 552−557.

(33) ChemAxon Ltd. http://www.chemaxon.com (accessed March 2007).

(34) 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.

(35) Shenk, J. S.; Westerhaus, M. O. Calibration the ISI Way. In *Near Infrared Spectroscopy: The Future Waves*; NIR Publications: Chischester, U.K., 1996; pp 198−202.

(36) Urbano Cuadrado, M.; Luque Ruiz, I.; Gómez-Nieto, M. A. A Steroids QSAR Approach Based on Approximate Similarities Measurements. *J. Chem. Inf. Model.* **2006**, *46*, 1678−1686.

(37) 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.

(38) Mannhold, R.; van de Waterbeemd, H. Substructure and whole molecule approaches for calculating log P. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 337−354.

CI700180M