Combining Similarity and Dissimilarity Measurements for the Development of QSAR Models Applied to the Prediction of Antiobesity Activity of Drugs

Irene Luque Ruiz, Manuel Urbano Cuadrado, and Miguel Ángel Gómez-Nieto

Abstract—In this paper we study different similarity based approaches for the development of QSAR model devoted to the prediction of activity of antiobesity drugs. Classical similarity approaches are compared regarding to dissimilarity models based on the consideration of the calculation of Euclidean distances between the nonisomorphic fragments extracted in the matching process.

Combining the classical similarity and dissimilarity approaches into a new similarity measure, the *Approximate Similarity* was also studied, and better results were obtained. The application of the proposed method to the development of quantitative structure-activity relationships (QSAR) has provided reliable tools for predicting of inhibitory activity of drugs. Acceptable results were obtained for the models presented here.

Keywords—Graph similarity, Nonisomorphic dissimilarity, Approximate similarity, Drugs activity prediction.

I. INTRODUCTION

GRAPH theory has been widely applied in Computational Chemistry. Chemists use 2D graph representations of chemical structures (called molecular graphs) in order to extract graph properties which are later related with chemical, physical and activity properties of the molecules represented using those graphs. Thus, several graph applications in the analysis and solutions of chemical problems have been carried out, namely: Quantitative Structure Activity/Property Relationships (QSAR/QSPR), query methods against large databases of chemical compounds, etc. [1],[2].

QSAR methodology seeks mathematical equations that correlate structural descriptors with activities of drugs as well as other pharmacological properties. This methodology shows a series of advantages, namely: non requirement of a deep theoretical knowledge of the receptor-drug system and the predictive ability achieved for a wide drug spectrum [3].

QSAR developments usually involve three stages: 1) first, descriptors, considered as chemical information vectors, are obtained from the data set (molecules) [4]; 2) use of mathematical regression techniques which establish formulas —usually multivariate expressions— that relate descriptors with activities or properties; 3) consideration of validation strategies in order to assess the predictive ability of QSAR equations according to analytical characteristics like accuracy, speed, robustness and reversibility [5].

One of the most widely used QSAR models are based on obtaining the space representation through the similarity calculation among the data set elements.

The basic idea underlying on similarity-based QSAR approaches was enunciated explicitly by Johnson and Maggiora [6], who state that "*molecules that are structurally similar likely will have similar properties*". Thus, 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.

Studies of similarity between chemical structures can be also overtaken using graphs. There are two stages involved in classical similarity calculations: 1) isomorphic subgraphs detection and extraction; and 2) similarity computation taking into account the number of isomorphic nodes and edges, that is, those nodes and edges common to the two matched graphs [7].

However, nonisomorphic fragments, which are not computed for the calculation of classical similarity measurements, also determine the properties and activities of chemical substances. Thus, QSAR models based on similarity approaches show high degeneracy because of information about nonisomorphic fragments is taken into account.

Therefore, in order to improve the accuracy and precision of chemical predictions, we propose the measurement of the similarity/dissimilarity between the nonisomorphic fragments extracted in the matching process, for the correction of classical similarity models.

I. Luque Ruiz is with Department of Computing and Numerical Analysis. University of Córdoba, Campus de Rabanales. Albert Einstein Building, E14071 Córdoba. Spain (e-mail: ma1lurui@uco.es).

M. Urbano Cuadrado is with Institute of Chemical Research of Catalonia ICIQ, Avinguda Països Catalans, 16. E-43007 Tarragona. Spain (e-mail: murbano@iciq.es)

M. A. Gómez-Nieto is with Department of Computing and Numerical Analysis, University of Córdoba, Campus de Rabanales, Albert Einstein Building, E14071 Córdoba, Spain (e-mail: mangel@uco.es).
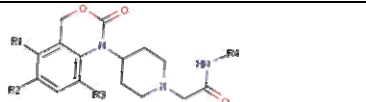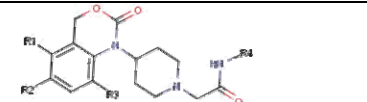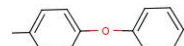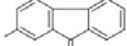
World Academy of Science, Engineering and Technology
International Journal of Pharmacological and Pharmaceutical Sciences
Vol:1, No:4, 2007
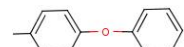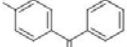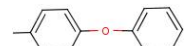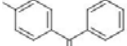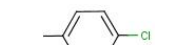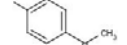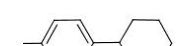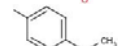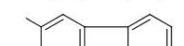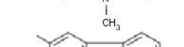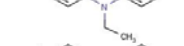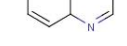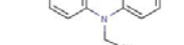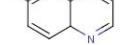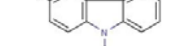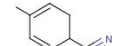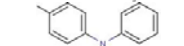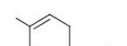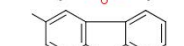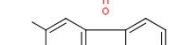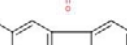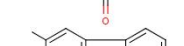
This dissimilarity value can be obtained through the calculation of molecular descriptors over the graphs (no necessarily connected) representing the nonisomorphic fragments, and a distances metric like Euclidean, Mahalanobis, and so on. The use of molecular descriptors allows the consideration of both size of nonisomorphic fragments and the type and nature of the nodes and edges (atoms and bonds).

Thus, a new and finer similarity measurement can be obtained, which overcomes disadvantages related to the non consideration of nonisomorphic subgraphs in the similarity calculation [8]. The corrected similarity measurement is called "*Approximate Similarity*" (*AS*), which merges isomorphic and nonisomorphic information into a more real similarity value since the difference between the subgraphs which do not form the isomorphism is employed for correcting classical similarity values.

This work has been organized as follows: after the introductory section, we describe the sample selected to apply the QSAR approach presented in this paper and the experimental procedure followed. Section 3 shows the experimental results obtained for the classical similarity, the use of nonisomorphic fragments and the combining of both in the AS similarity measurements. Finally, conclusions are given in section 4.

TABLE I
SAMPLE OF 30 CHEMICAL COMPOUNDS AND THEIR ANTIOBESITY ACTIVITY VALUES USED IN THE TEST

	R1	R2	R3	R4	pIC$_{50}$		R1	R2	R3	R4	pIC$_{50}$
1	H	H	H	(structure)	1.30	16	H	Cl	H	(structure)	1.84
2	H	CH$_3$	H	(structure)	2.02	17	H	H	CH$_3$	(structure)	1.60
3	H	Cl	H	(structure)	1.78	18	H	Cl	H	(structure)	1.05
4	H	H	H	(structure)	2.48	19	H	H	H	(structure)	1.24
5	H	Cl	H	(structure)	2.05	20	H	Cl	H	(structure)	1.32
6	H	H	H	(structure)	0.98	21	H	CH$_3$	H	(structure)	1.41
7	Cl	H	H	(structure)	1.74	22	H	F	H	(structure)	0.89
8	H	Cl	H	(structure)	2.00	23	H	Cl	H	(structure)	2.17
9	H	H	CH$_3$	(structure)	1.70	24	H	H	H	(structure)	2.12
10	H	H	OCH$_3$	(structure)	2.88	25	H	Cl	H	(structure)	2.30
11	H	H	OCH$_3$	(structure)	1.75	26	H	H	H	(structure)	0.88
12	H	H	H	(structure)	1.37	27	H	H	H	(structure)	2.14
13	H	H	CH$_3$	(structure)	1.70	28	H	Cl	H	(structure)	0.90
14	H	Cl	H	(structure)	1.40	29	H	H	H	(structure)	0.94
15	H	CH$_3$	H	(structure)	1.94	30	H	H	CH$_3$	(structure)	1.48

World Academy of Science, Engineering and Technology
International Journal of Pharmacological and Pharmaceutical Sciences
Vol:1, No:4, 2007

II. QSAR MODEL FOR THE PREDICTION OF ANTIOBESITY ACTIVITY OF DRUGS

A. Sample of Antiobesity Drugs

In this paper we try to analyse the capacity of different similarity approaches in the development of QSAR models for the prediction activity of NPY Y5 antagonist given their potential antiobese agents [9].

It is widely accepted that obesity influences on many kinds of diseases and disorders, namely: respiratory, musculoskeletal, gastrointestinal, cardiovascular, etc. So, the development of effective and safe antiobesity drugs has a high interest to pharmacological chemists.

The data set selected were 30 benzoxazinone derivates selected as the chemical space to be modelled. These compounds have shown good antiobese activity in recent studies [9]. Table I shows the data set 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.

B. Data Modeling and Calculation

Graph structures were built using *MarvinSketch* software [10]. Structural isomorphism for the calculation of classical similarity measurements was obtained using fingerprint and graph matching approaches.

Fingerprints were generated by *generfp* (default options) of Jchem [10]. A fingerprint is a binary array of a preset size which represents structural properties of the molecular graphs. Different kinds of fingerprints have been proposed depending on the structural elements to be represented and on the array size. In a general way, fingerprint construction consists of a series of steps, namely: a) generation of the molecular graph for each element of the data set; b) obtaining of the subgraphs showing size from 1 to m (often lower than 9) for each graph; c) extraction of preset pattern substructures in some fingerprint kinds; d) assignation of a binary representation and position of each path and pattern presented in the data set; e) and finally, the fingerprint construction.

So, fingerprints can be considered as data structures which do not require great computational costs for their handling and store greater structural information than that shown by the chemical graph. Fingerprint similarity values are calculated through the computation of Boolean operations among the bits of the fingerprints.

Furthermore, structural isomorphism using graph matching was calculated for all the pairs of data set molecules by using an algorithm developed by the authors [11].

MCS (maximum common substructure) was the isomorphic fragment considered. Isomorphic and nonisomorphic fragments were retained for descriptor and similarity calculation.

For the calculation of dissimilarity measurements between nonisomorphic fragments Hyper-Wiener descriptor was considered. As expression 1 shows, Hyper-Wiener index (*WW*) is derived from Wiener index (see expression 2) but also considers a quadratic term of the distance contribution of the graph nodes.

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

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

Hyper-Wiener (also Wiener) index is calculated thought the graph distance matrix. In this matrix the distance between two nodes connected is equal to 1, therefore the characteristics between different types of nodes is not taken into account.

In order to consider the nature of the nodes and edges of the molecular graphs we have used weighted distance matrix for the calculation of Hyper-Wiener index.

In weighted distance matrices distance equal to 1 between nodes *i* and *j* are replaced by bond length between the atoms *i* and *j* relative to the distance to the bond *C-C*. Software developed by the authors was used for this calculation.

In addition, dissimilarity matrices (Euclidean-based distances) were employed for the measurement of nonisomorphic fragments. Then these matrices were used for QSAR model construction.

C. Multivariate Regression Analysis

If an *N* by *N* similarity matrix is built using *N* compounds, this matrix can be employed to develop multivariate QSAR approaches. Each matrix element *(i,j)* provides the similarity between the compounds *i* and *j* and it shows the same value as the element *(j,i)*. The diagonal of the matrix is equal to 1.

From the point of view of multivariate regression, the matrix is considered a set of *N* objects (rows) characterized by *N* variables (columns). Thus, an object is a given compound described by a serie of global variables which accounts for the similarity between the compound and a reference compound. PLS was employed as the multivariate regression technique [12].

III. EXPERIMENTAL RESULTS

A. Classical Similarity Analysis

First, constitutional and fingerprint-based similarity matrices were built in order to analyze the results obtained by means of considering only isomorphic information.

Table II shows the statistical results and the number of property-descriptor outliers obtained in LOO processes.

The outlier study was carried out by setting a cut-off 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

World Academy of Science, Engineering and Technology
International Journal of Pharmacological and Pharmaceutical Sciences
Vol:1, No:4, 2007

anomalous behaviour with regard to the rest of the data set.

A $T_{cut\text{-}off}$ set to 2.5 is often used in multivariate models. As can be observed in Table II, in spite of obtaining good correlations once outliers had been removed, the number of outliers was excessive (it can not be greater than 10 % of the data set size).

TABLE II
STATISTICAL RESULTS FOR CONSTITUTIONAL AND FINGERPRINTS BASED
SIMILARITY APPROACH

	Q^2	Slope	Bias	SECV	Outlier
Constitutional	0.73	0.87	0.20	0.22	6
Fingerprints	0.75	0.79	0.33	0.28	5

Fig. 1 shows the two first principal components plots obtained with PCA analysis for Constitutional similarity approach. The six outliers detected with this approach can be observed clearly.

Molecules 10 and 11 are sited alone in the lower quadrant. Molecule 10 has the maximum activity value of de data set, being difficult to modelled, and molecule 11 contains a fragment which does not exist in another data set molecule. Similar interpretation can be given for molecule 4.

On the other hand, molecules 22 and 23 are very similar but they show a very different activity value. The difference between the chemical properties of –F and –Cl substituents is not enough to interpret the high activity difference between both molecules. A close interpretation can be given for the outlier corresponding to molecule 27 when its structure and activity value are compared with molecule 12. Both molecules are very similar, however a high activity difference exist.
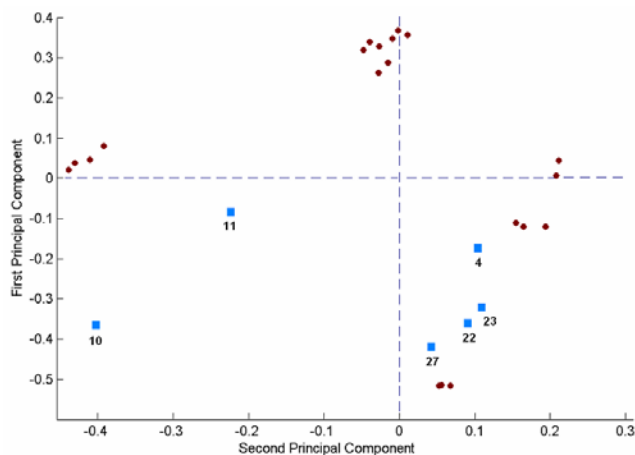

Fig. 1 PC1 vs. PC2 for Constitutional similarity approach

In Fig. 2 PC1 vs. PC2 is plotted for the fingerprints-based approach. As we observed, the outliers 4, 7, 11, 22, 23 are also detected and not modelled.

This behaviour of classical similarity methods could be due to not considering directly the nonisomorphic data in modelling, thus leading to strong problems related to the *cliff* phenomenon.
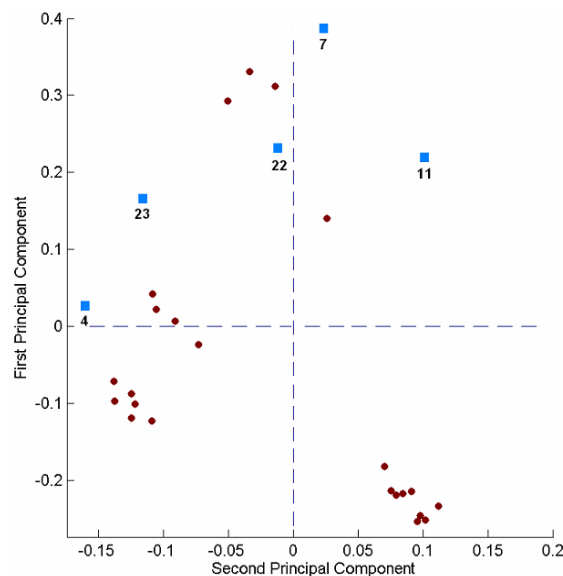

Fig. 2 PC1 vs. PC2 for Constitutional similarity approach

B. Nonisomorphic Fragments Behavior

Several works have been proposed by taking into account differences between nonisomorphic fragments extracted from matching processes with the aim of modelling different data sets [2],[3].

Dissimilarity measures can be obtained by means of using an appropriate descriptor over molecular graphs (not necessarily full connected) that correspond to nonisomorphic fragments.

If the matching algorithm proposed by authors [42] is applied to the molecules of a given data set, 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*. Dissimilarity or distance value can be obtained as follows:

$$\Gamma_{A,B} = \frac{[TD^2(NIF_A) + TD^2(NIF_B)]^{0.5}}{TD(A) \times TD(B)} \qquad (3)$$

where: *TD(A)* and *TD(B)* represent the descriptors computed over the molecules *A* and *B*, respectively, and *TD(NIF_A)* and *TD(NIF_B)* show the invariant value for nonisomorphic structures of *A* and *B*. Thus, $\Gamma_{A,B}$ is a dissimilarity value which accounts for the nonisomorphic fragments of the matched molecules.

When expression 3 was used with Hyper-Wiener index and weighted distance matrix, the statistical results were as follows: Q^2 = 0.77, **Slope** = 1.02, **Bias** = 0.33, **SECV** = 0.24, **Outliers** = 3.

Therefore, dissimilarity approach based on nonisomorphic fragments shows a better correlation coefficient than classical similarity model, also reducing the number of outliers and standard error.

World Academy of Science, Engineering and Technology
International Journal of Pharmacological and Pharmaceutical Sciences
Vol:1, No:4, 2007

We also tested expression 3 with both non-weighted distance matrices and Wiener index and not noteworthy differences were observed. Fig. 3 shows PC1 vs. PC2 for dissimilarity approach of expression 3.

As observed in Fig. 3, molecule 22 is detected again as outlier, although the other outliers detected by classical similarity approach are modelled. However, although the number of outlier is reduced to only 3, molecules 1 and 6 can not be modelled.
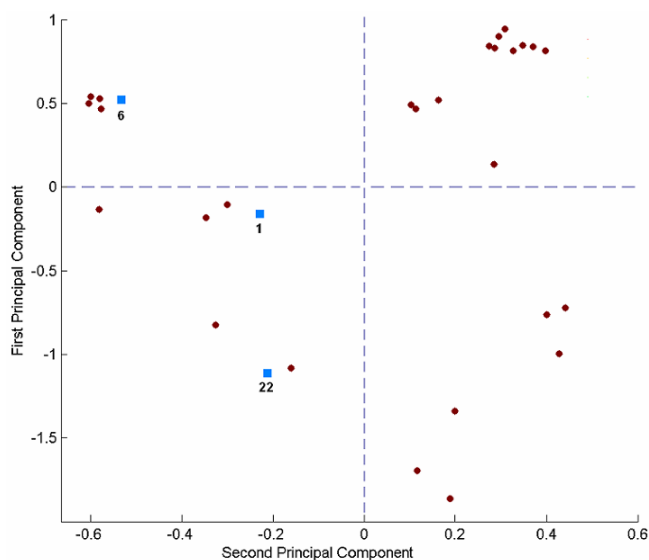

Fig. 3 PC1 vs. PC2 for dissimilarity approach based on nonisomorphic fragments

From the above commented results, we can conclude because of the data set characteristics: a) weighted and non-weighted distance matrices show similar behavior because few heteroatom are present in the non isomorphic fragments, b) Wiener index shown worst results than Hyper-Wiener because no very large nonisomorphic fragments are detected and, therefore the consideration of squared terms (see expression 1) by means of the Hyper-Wiener index involves a refinement of the nonisomorphic structure-based influence on the distance measure, and c) the characteristics of nonisomorphic fragments is a good approximation to the development of a prediction model, even more accuracy than the models based on isomorphic fragments.

C. Combining QSAR Approaches

For the different approaches above described a set of outliers were detected. We can classify outliers into two types, namely: a) those which do not belong to the chemical space to be modelled (structural or activity space), thus requiring new objects to cover the non well defined chemical regions; and b) outliers characterized by the *cliff* phenomenon [15], which derives from great activity variations resulting from small structural changes. So, solutions to the *cliff* outliers are difficult to be solved.

The fact of detecting molecules 22 and 23 as outliers is owing to the *cliff* problem since these two molecules are extremely similar and show quite different activities, as can be observed in Table I and Fig. 1. On the other hand, other outliers show problems related to the incorrect definition of the activity space to predict. We can observe that there are some activity value intervals only defined by one object (one compound), thus leading to poor predictions for these molecules.

As stated in the Introduction 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 [8], [13], [14] we have shown the usefulness of the Approximate Similarity (*AS*) concept to develop QSAR models. Approximate similarity is based on correcting classical 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, as we described in expression 3.

Thus, the Approximate Similarity (*AS*) is defined as follows:

$$AS_{A,B} = f(S_{A,B}, \Gamma_{A,B}, w_\Gamma) \tag{4}$$

where: $S_{A,B}$ is a classical similarity measurement (constitutional or fingerprint-based); $\Gamma_{A,B}$ is the distance or dissimilarity measurement obtained through the nonisomorphic fragments as stated by expression 3, and w_Γ is a weighting factor which adjusts the contribution of the nonisomorphic fragments on the similarity measurement.

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, *AS* expression was as follows:

$$AS_{A,B} = S_{A,B} - \left[\Gamma_{A,B} - abs\left(\frac{TD(MCS)}{TD(A)} - \frac{TD(MCS)}{TD(B)} \right) \right] \tag{5}$$

where: *TD(MCS)* is the value of the Hyper-Wiener index corresponding to the isomorphic fragment (maximum common subgraph) extracted from the matching of *A* and *B* molecules, 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.

Fig. 4 shows PC1 vs. PC2 plot corresponding with the model built using approximate similarity. Three outliers are detected. As observed, molecule 22 is detected again as an outlier. The extremely low activity value of this molecule regarding to similar molecules of the data set makes very difficult model it.

World Academy of Science, Engineering and Technology
International Journal of Pharmacological and Pharmaceutical Sciences
Vol:1, No:4, 2007

On the other hand, the group composed by molecules 1, 2 and 3 (rounded in Fig. 4) is also difficult to model. Molecule 2 is detected as an outlier, while using dissimilarity approach molecule 1 is detected. These three molecules show a very close structure but very different activity value.
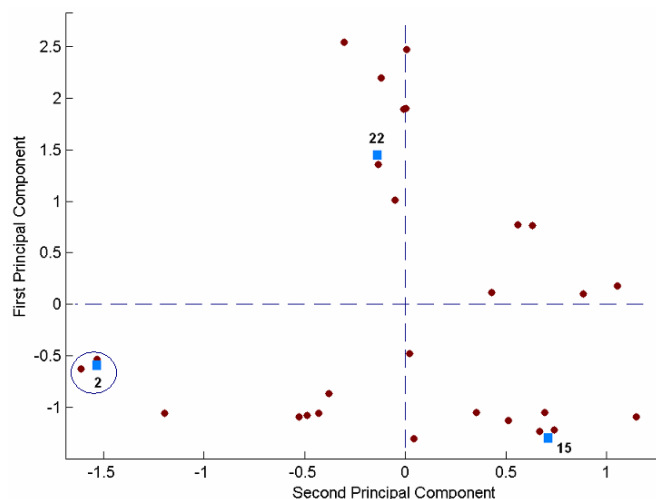

Fig. 4 PC1 vs. PC2 for AS approach

Fig. 5 shows the representation of the differences between experimental and predicted pIC_{50} values obtained with expression (5) for the data set without the outliers' consideration. Statistical characterization of the prediction capacity was as follows: $Q^2 = 0.88$, *slope = 1.06* and ***bias = 0.09***, ***SECV = 0.18***

Therefore, accuracy and precision of the predictions carried out were significantly improved, thus leading to useful uncertainty reductions for the computer-aided drug development.

IV. CONCLUSION AND REMARKS

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 shown an anomalous behaviour owing to the *cliff* phenomenon or to problems related to the chemical space definition.

The lowest number of outliers was obtained for nonisomorphic matrices: 3 compounds shown anomalous deviations, this quantity is lower than the number of outliers accepted by chemometric community as cut-off for the development of predictive models (15 % of the data set size).

Combining of dissimilarity measurements calculated over nonisomorphic fragments and classical similarity measurements gave an excellent result.

Thus, we can conclude that the QSAR development requires most of the times 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 modelled.

It is interesting to remark the fact of employing 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.
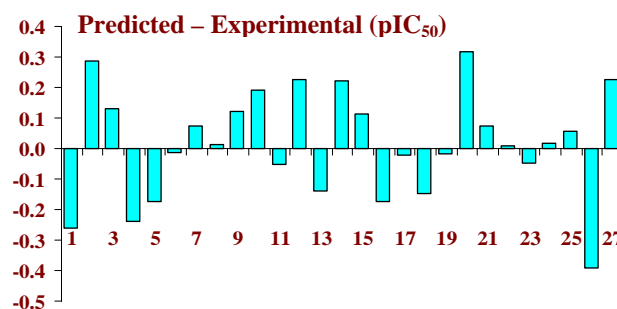

Fig. 5 Differences between Predicted and Experimental pIC_{50} values obtained for the prediction model of eq. (5)

REFERENCES

[1] Rouvray, D.H.; Balaban, A.T. Chemical Applications of Graph Theory. Applications of Graph Theory. Wilson, R.J.; Beineke, L.W. (Eds.). Academic Press. 1979, 177-221.
[2] 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. The Netherlands. 1999, 59-167.
[3] van de Waterbeemd, H.; Gifford, E. ADMET in silico modelling: towards prediction paradise? *Nat. Rev. Drug Discov.* 2003, 2, 192–204.
[4] Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O. A stepwise approach for defining the applicability domain of SAR and QSAR models. *J. Chem. Inf. Model.* 2005, 45, 839–849.
[5] Nikolova, N. and Jaworska, J. Approaches to measure chemical similarity – a review. QSAR Comb. Sci. 2004, 22, 1006–1026.
[6] Johnson, M.A.; Maggiora, G.M. eds. *Concepts and Applications of Molecular Similarity*. John Wiley, 1990.
[7] Willett, P. Chemical Similarity Searching. J. Chem. Inf. Comput. Sci. 1998, 38, 983-996.
[8] Urbano Cuadrado, M.; Luque Ruiz, I.; Gómez-Nieto, M.A. A New Quantitative Structure-Property Relationship Based on Topological Distances on Nonisomorphic Subgraphs. In Lectures Series on Computer and Computational Sciences: Advances in Computational Methods in Sciences and Engineering. Brill Academic Publisher, 2005. 135-138.
[9] Deswal, S.; Roy, N. Quantitative structure activity relationship of benzoxazinone derivatives as neuropeptide Y Y5 receptor antagonists. *European Journal of Medicinal Chemistry*. 2006, 41 552–557.
[10] ChemAxon Ltd. http://www.chemaxon.com. Last acceded March, 2007.
[11] 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.
[12] Wold, S.; Sjostrom, M.; Eriksson, L. PLS-Regression: A Basic Tool of Chemometrics, Chemom. Intell. Lab. Syst. 2001, 58, 109-130.
[13] 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.
[14] 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.
[15] Maggiora, G. F. On Outliers and Activity Cliffss - Why QSAR Often Disappoints. *J. Chem. Inf. Model.* 2006, 46, 1535-1535.