International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:1, No:2, 2007

New Graph Similarity Measurements based on Isomorphic and Nonisomorphic Data Fusion and their Use in the Prediction of the Pharmacological Behavior of Drugs

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

***Abstract*—**New graph similarity methods have been proposed in this work with the aim to refining the chemical information extracted from molecules matching. For this purpose, data fusion of the isomorphic and nonisomorphic subgraphs into a new similarity measure, the Approximate Similarity, was carried out by several approaches. The application of the proposed method to the development of quantitative structure-activity relationships (QSAR) has provided reliable tools for predicting several pharmacological parameters: binding of steroids to the globulin-corticosteroid receptor, the activity of benzodiazepine receptor compounds, and the blood brain barrier permeability. Acceptable results were obtained for the models presented here.

***Keywords*—**Graph similarity, Nonisomorphic dissimilarity, Approximate similarity, Drug activity prediction.

I. Introduction

SINCE the development of the Graph Theory, chemists have shown a great interest in the representation of 2D chemical structures by means of graphs. Arisen from this representation, several applications of graphs in the analysis and solution 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].

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 (those nodes and edges common to the two matched graphs) [3].

The use of graph similarity measurements for the development of QSAR methods is a work topic aimed at obtaining tools for predicting the pharmacological activity of drugs. Based on the “structurally similar molecules show similar properties and biological activities” chemical principle [4], graph similarity provides QSAR tools characterized by simplicity and fastness.

In order to improve the accuracy and precision of chemical predictions, we propose a new graph similarity measurement, the Approximate Similarity (*AS*), which overcomes disadvantages related to the non consideration of nonisomorphic subgraphs in the similarity calculation [5]. Thus, the *AS* value merges isomorphic and non isomorphic 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.

Other two developments also aimed at improving the predictive ability of similarity measurements are presented. First, we propose the use of topological invariants for describing both the isomorphic subgraphs and the complete graphs employed in the classical similarity calculation. Thus, the type and nature of the nodes and edges (atoms and bonds) are considered instead of the simple approach which takes into account only the number of both nodes and edges.

Second, we have developed a new isomorphism detection method which also computes the nodes and edges bridges between the isomorphic and nonisomorphic subgraphs, in addition to common substructures. This new method allows us to consider the position and nature of substituents, which are keys for the pharmacological activity of chemical structures.

This work has been organized as follows: after the introductory section, a general description of the approximate similarity concept is given in section 2. Section 3 shows the building and test of several AS-QSAR models which pursue the prediction of drug parameters in several families. Finally, conclusions are given in section 4.

Manuscript received January 20, 2007. This work was supported by the Comisión Interministerial de Ciencia y Tecnología (CiCyT) and FEDER (Project: TIN2006-02071).

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 (email: 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).

II. The Approximate Similarity

A. Similarity and Distance Concepts in Graph Matching

Given two graphs G_A and G_B of size (number of nodes and edges) A and B, respectively, which represent, as shown in Fig. 1, the molecules M_A and M_B, we define $I_{A,B}$ as the isomorphism present between these graphs, and NIF_A and NIF_B as the non common parts (nonisomorphic subgraphs) between G_A and G_B. The structural similarity can be calculated as follows:

$$S_{A,B} = f(I_{A,B}, A, B) \tag{1}$$

where f is a function which matches $S_{A,B}$ and $I_{A,B}$ taking into account the size of graphs G_A and G_B.

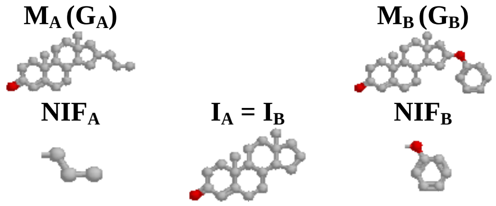


Fig. 1 Description of molecules M_A and M_B by graphs G_A and G_B. These molecules present the isomorphism $I_{A,B}= I_A = I_B$. NIF_A and NIF_B represent the subgraphs of G_A and G_B, respectively, which do not form $I_{A,B}$

The similarity $S_{A,B}$ is a value in the interval [0,1] which gives the similarity between the graphs G_A and G_B and, then, the closer to 1 the $S_{A,B}$ shows, the higher the similarity between the molecules M_A and M_B is. Thus, different similarity values are obtained depending on both the method employed for calculating the isomorphism and the *f()* similarity function (similarity index considered). Regarding the isomorphism calculation, *MCES* (Maximum Common Edges Subgraph), *MCS* (Maximum Common Subgraph) or *AMCS* (All Maximum Common Subgraphs) approaches, in addition to the methods based on transforming graphs into fingerprints, are the commonest methodologies [6]. And regarding to the similarity function, there are several similarity indexes summarized in literature whose difference lies in the function, namely: Tanimoto, Cosine, Simpson, Raymond, etc. [2].

Since our proposal is also to consider distances between the subgraphs that do not form the isomorphism $I_{A,B}$, the structural difference $\Gamma_{A,B}$ (dissimilarity or distance) between two molecular graphs G_A and G_B is calculated as follows:

$$\begin{aligned}\Gamma_{A,B} &= g[td(G_A, I_{A,B}), td(G_B, I_{A.B})] = \\ &= g[td(NIF_A), td(NIF_B)]\end{aligned} \tag{2}$$

where $I_{A,B}$ has the equal meaning to that shown in expression (1); NIF_A=G_A-$I_{A,B}$ and NIF_B=G_B-$I_{A,B}$ represent the subgraphs of G_A and G_B, respectively, that do not form the isomorphism $I_{A,B}$; *g()* is a function aimed to obtaining a distance value (e.g. Euclidean, Mahalanobis, etc.) between *td*(NIF_A) and *td*(NIF_B); and *td* is a topological descriptor which describe the noncommon subgraphs, namely: Wiener (W), Hyper Wiener (WW) and so on indexes. Contrary to similarity, higher the $\Gamma_{A,B}$ shows, higher the dissimilarity between the molecules M_A and M_B is.

B. Correction of the Structural Similarity: The Approximate Similarity

With the aim of defining a new similarity measurement which takes into account both the classical similarity and the nonisomorphic distance, the Approximate Similarity (*AS*) is defined as follows:

$$AS_{A,B} = f(S_{A,B}, \Gamma_{A,B}, w_\Gamma) \tag{3}$$

where $S_{A,B}$ and $\Gamma_{A,B}$ are the similarity and dissimilarity defined in equations (1) and (2), respectively; and w_Γ is a weighting factor which adjusts the distance contribution in the approximate similarity calculation.

Thus, chemical similarity achieved by the *AS* approach will be more accurate due to the consideration of the difference between the noncommon substructures of the matched molecules, that most time are responsible for their properties / activities.

C. Multivariate AS Predictive Spaces

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

From the point of view of multivariate regression, the *AS* 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 [7].

III. AS-QSAR Models

Several QSAR models which are able to match the *AS* values of a drug with its pharmacological activity have been built. For this purpose, chemical data sets were splited into training and test subsets; the former was employed for building the models (full cross validation was the training strategy), whereas the latter was used for externally validating the predictive ability achieved.

Three AS-QSAR models for three pharmacological activities, respectively, are presented below.

A. Model for Predicting the Steroid Binding to the Corticosteroid-Binding Globulin Receptor

The thirty classical steroids considered as the benchmark for testing structure activity-relationships [8] was the first chemical family to be modeled by *AS* matrixes. *AS* measurements were computed as follows:

$$AS_{A,B} = S_{A,B} - w_{\Gamma} \times \overline{\Gamma}_{A,B} \qquad (4)$$

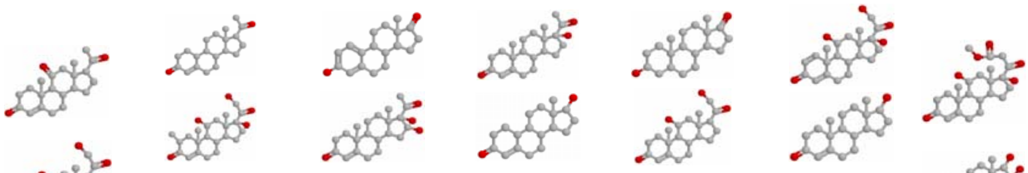
where $\overline{\Gamma}_{A,B}$ is a scaled value of the structural dissimilarity $\Gamma_{A,B}$ defined in equation (2). Similarities were calculated using the cosine index, whereas dissimilarities were computed using both the Euclidean distance and the Wiener invariant (computed the latter over the weighted distance matrix of the nonisomorphic subgraphs).

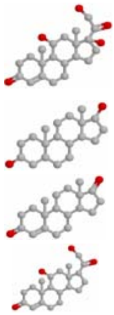
There is a need to optimize the weight w_{Γ} for combining structural similarities and distances. An excessive significance of distances in *AS* calculation can add random information and, then, no a deterministic part to the total predictive ability. In this study, the factor w_{Γ} was moved from 1.0 to 0.1 and important variations were obtained. Different models were trained using 1-21 steroids and then tested using the remaining 22-30. Table I shows the statistical parameters obtained in the training and test of the models built using the 30 steroids also shown in Table I.

The higher the Q^2 and *r* parameters are, the greater the predictive ability is (minimal error). *Slope* and *bias* must be close to 1.00 and 0.00, respectively. As can be observed, the best model was developed when *AS* matrixes were computed using the factor w_{Γ} set at 0.3. As can be observed in Table I this model also shows a higher predictive ability than the models using classical similarity and distance separately.

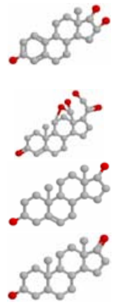
Moreover, the results obtained compare reasonably well with other recent methods based on 3D-QSAR methods [9]. It should be stressed that our model is based on topological measurements and, then, it is a simpler method than other approaches.

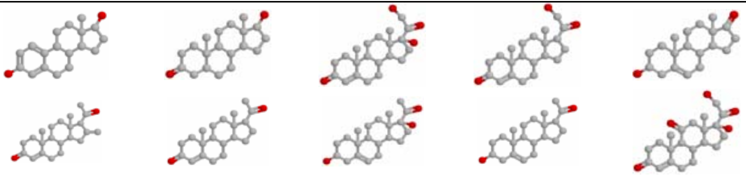
TABLE I
STATISTICAL PARAMETERS OBTAINED IN THE TRAINING AND TEST STAGES OF SEVERAL AS-QSAR MODEL BUILT USING THE 30 STEROIDS ABOVE REPRESENTED





	Training				**Test**			
Descriptor	**Q^2**	**Error**	**Slope**	**Bias**	**r**	**Error**	**Slope**	**Bias**
Similarity	0.80	0.53	0.97	0.15	0.68	0.45	0.95	0.20
Distance	0.60	0.74	0.98	0.14	0.61	0.55	0.54	3.29
***AS* (w_{Γ} =1.0)**	0.71	0.66	0.85	0.90	0.28	0.78	-0.39	9.70
***AS* (w_{Γ} =0.5)**	0.82	0.51	0.97	0.17	0.77	0.39	0.90	0.58
***AS* (w_{Γ} = 0.3)**	0.84	0.47	0.97	0.15	0.82	0.39	1.03	-0.45
***AS* (w_{Γ} = 0.1)**	0.77	0.57	0.97	0.17	0.65	0.44	0.79	1.38





B. Model for Predicting the Activity of Benzodiazepine Receptor Ligands

The GABAA/benzodiazepine receptor (GABAA/BzR) forms a chloride ion (Cl¯) selective channel. Its function is initiated by the binding of γ-aminobutyric acid (GABA), considered as the principal inhibitory neurotransmitter of the central nervous system. GABAA/BzR agonists (anxiolytic, anticonvulsant, and sedative effects), inverse agonists (anxiogenic, stimulant, and convulsant effects) or antagonists (null efficacy) are recognized compounds which bond to GABA/BzR and enhance, diminish or block the Cl¯ channel, respectively. We have developed *AS-QSAR* models for predicting the activity of 58 compounds which bond to GABA/BzR.

With the aim of improving several characteristics of the *AS* concept, we propose the *AS* calculation as follows:

$$AS_{A,B} = S^{I}_{A,B} * \left(1 - abs\frac{TD(NIF_A) - TD(NIF_B)}{TD(A) + TD(B)}\right) \qquad (5)$$

where $S^{I}_{A,B}$, the invariant-based similarity, which uses topological descriptors in similarity calculation instead of the number of nodes and edges. Thus, $S^{I}_{A,B}$ should refine the extracted chemical information and, in turn, improve the structure-activity relationships —type and ratio of intramolecular bonds are employed—.

On other hand, $S^{I}_{A,B}$ is corrected taking into consideration the size and nature of the molecules *A* and *B* (values of *TD(A)* and *TD(B)*), and the dissimilarity of the nonisomorphic parts (values of $TD(NIF_A)$ and $TD(NIF_B)$). The greater difference between $TD(NIF_A)$ and $TD(NIF_B)$ is the greater similarity correction, having this factor values 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.

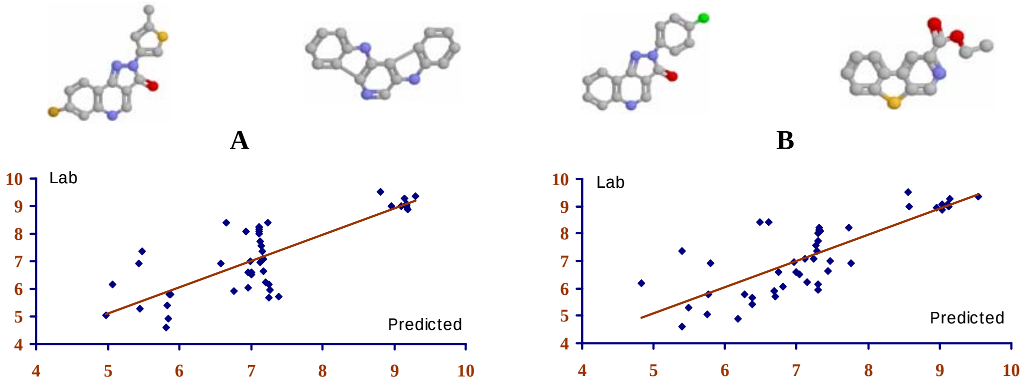
Fig. 2 (A) and (B) shows the predicted vs lab activities plots obtained in the training stage (using 49 of the 58 compounds) for the *AS* matrixes built using eq. 4 — w_Γ set at 0.3— and eq. 5, respectively. Test stages carried out with the remaining 9 compounds gave the following results:

Eq.4 Model: r = **0.75**; *Error* = **0.98**;
slope = **0.74**; *intercept* = **2.03**
Eq.5 Model: r = **0.79**; *Error* = **0.8**2;
slope = **1.00**; *intercept* = **0.00**

Taking into consideration the above results and Fig. 2, the new contributions to the *AS* approach are responsible for the better results achieved with eq. 5. In addition, results again compare reasonably well with those obtained with more complex approaches [10].

C. Model for Predicting the Blood Brain Barrier Permeability of Drugs

Since the previous AS-QSAR models have been developed for homogenous chemical families, the use of *AS* for predicting the blood brain barrier permeability (BBBP) —expressed as the logBB— of 130 compounds belonging to very different chemical families was here pursued. BBBP informs about the physical barrier strength stopping substances from traveling into the central nervous system.



*Q²=***0.60***, SECV=***0.91***, Slope=***0.95***, Intercept=***0.33** *Q²=***0.64***, SECV=***0.83***, Slope=***0.97***, Intercept=***0.30**

Fig. 2 Lab *vs*. predicted activity plots and statistical parameters of the training stages carried out using (A) eq. 4 and (B) eq. 5. Data set compounds have similar structures to the ones above shown

TABLE II
STATISTICAL PARAMETERS OBTAINED IN THE TRAINING AND TEST OF THE *AS* MODELS BUILT USING EQ. 6 AND THE WIENER (W) AND HYPERWIENER (WW) INVARIANTS

	Training				**Test**			
Approach	**Q²**	**Error**	**Slope**	**Bias**	**r²**	**Error**	**Slope**	**Bias**
W index	0.73	0.31	0.98	-0.01	0.91	0.35	1.30	0.03
WW index	0.71	0.32	0.96	-0.01	0.91	0.34	1.30	0.04

An important new contribution to the *AS* concept was proposed in the building of this model: the Extended Maximum Common Subgraph (*EMCS*), a new isomorphism detection which also accounts for the nodes and edges which are bridges between the isomorphic and nonisomorphic fragments.

In this way, a nonsymmetric matrix where each element stores the *EMCS(i,j)* of the molecule *i* compared with the molecule *j* is obtained for a given dataset.

Using this matrix, *EMCS(i,j)* topological descriptors can be obtained and combined with similarities in order to obtain predictive spaces which reflect dissimilarities of the substituent positions. *AS* equation was as follows:

$$AS_{A,B} = S^{I}_{A,B} - abs\left(\frac{EMCS_{A,B}}{TD(A)} - \frac{EMCS_{B,A}}{TD(B)}\right) + \left(1 - abs\frac{TD(NIF_A) - TD(NIF_B)}{TD(A) + TD(B)}\right) \tag{6}$$

The first component is the invariant-based similarity and reflects the nucleus similarity regarding to the size and nature of molecules *A* and *B*. The second component is the term which uses the information about the position and nature of substituents. In this way, $EMCS_{A,\mathrm{B}}$ and $EMCS_{B,A}$ relative values regarding to *TD(A)* and *TD(B)*, respectively, are employed to measure a dissimilarity measurement of substituent nature and position.

Since the second term is a dissimilarity value, its contribution to the approximate similarity values is negative. It is the first time this information is employed. Finally, the last term adds a correcting factor proportional to the noncommon graphs dissimilarity (NIF_A and NIF_B).

The data set composed by 130 molecules was divided into the training (105) and test (25) sets. Table II shows the statistical characterization of the building and validation stages. Two topological invariants (Wiener and HyperWiener) were employed for isomorphic and nonisomorphic substructures, but similar results were obtained.

The difference of the calculation of both indexes lies on the analysis of the weighted distance matrix. Wiener index is computed as the half-sum of all the elements of the weighted distance matrix, while the HyperWiener descriptor uses quadratic values of the matrix elements in addition to the half-sum of these elements. Differentiation achieved by the Wiener graph description was appropriate for obtaining good results, but the consideration of HyperWiener indexes did not add noise to the predictive matrixes.

IV. Remarks

Methods presented in this work have achieved the aim of refining the classical similarity measure by means of considering differences or dissimilarities between the subgraphs which do not form the isomorphism between two molecular graphs.

Thus, a new similarity approach (Approximate similarity) has been employed for developing QSAR models for the prediction of several activities of drugs, namely: 1) binding of steroids to the globulin-corticosteroid receptor, (2) the activity of benzodiazepine receptor compounds, and (3) the blood brain barrier permeability. The *AS* approach was applied to homogeneous and heterogeneous data sets. For the latter, the consideration of Extended Maximum Common Subgraphs as a new isomorphism measurement allowed developing valuable prediction models.

Moreover, when our results were compared with other more complex QSAR approaches summarized in literature, better and similar statistical values were obtained. Robustness of models has also been tested by external validation, showing acceptable results.

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] Willett, P. Chemical Similarity Searching. J. Chem. Inf. Comput. Sci. 1998, 38, 983-996.

[4] Downs, G.M.; Barnard, J.M. Clustering and Their Uses in Computational Chemistry. In Reviews in Computational Chemistry. Lipkowitz, K.B., Boyd, D.B. (Eds.) Wiley-VCH. New York. 2003, 18, 1-39.

[5] Urbano Cuadrado, M.; Luque Ruiz, I.; Gómez-Nieto, M.A. A New Quantitative Structure-Property Relationship Based on Topological Distances on Non-isomorphic Subgraphs. In Lectures Series on Computer and Computational Sciences: Advances in Computational Methods in Sciences and Engineering. Brill Academic Publisher, 2005. 135-138.

[6] 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.

[7] Wold, S.; Sjostrom, M.; Eriksson, L. PLS-Regression: A Basic Tool of Chemometrics, Chemom. Intell. Lab. Syst. 2001, 58, 109-130.

[8] Silverman, B.D. The Thirty-one Benchmark Steroids Revisited: Comparative Molecular Moment Analysis (CoMMA) with Principal Component Regression. Quant. Struct.-Act. Relat. 2000, 19, 237-246.

[9] Coats, E.A. The CoMFA Steroids as a Benchmark Dataset for Development of 3D QSAR Methods. In 3D QSAR in Drug Design. Kubinyi, H., Folkers, G., Martin, Y.C. (Eds.). Kluwer/Escom. Dordrecht. 1998, 199-213.

[10] 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.