

# A new quantitative structure–property relationship approach using dissimilarity measurements based on topological distances of non-isomorphic subgraphs

Manuel Urbano-Cuadrado · Irene Luque Ruiz · Miguel Ángel Gómez-Nieto

Received: 30 March 2007 / Accepted: 30 September 2008 / Published online: 9 July 2009  
© Springer Science+Business Media, LLC 2009

**Abstract** A Quantitative Structure–Property Relationship model has been developed using a new method proposed in this paper, which is aimed at overcoming disadvantages related to the use of similarity calculations in quantitative approaches. The method uses the concept of topological descriptor but applied to non-isomorphic subgraphs. A symmetrical matrix comprising Euclidean distances according to differences between the non-isomorphic subgraphs is built. This symmetrical matrix is used as input of Partial Least Squares Regression processes for predicting sublimation enthalpies of Polychlorinated Biphenyls. Statistical results ( $R^2$  in full cross validation, Standard Error in Cross Validation, slope and bias) of our model were obtained and compared with those from the use of similarity values, univariate topological descriptors and literature approaches.

**Keywords** Graph theory · Topological descriptors · Similarity and distance · QSPR · PLSR

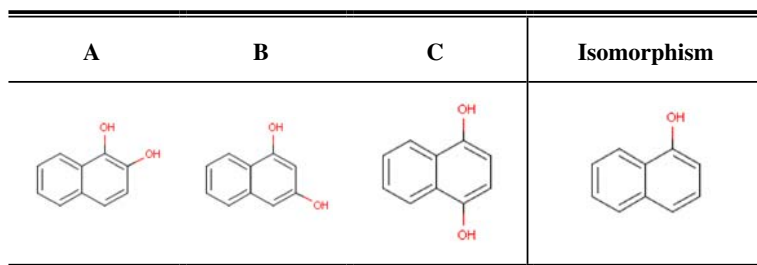
## 1 Introduction

Thanks to advances in computers, abstract models that are representations of scientific or technical problems or phenomena can be employed to derive predictive tools. Thus, the role of data explanation is expanded to place the role of prediction.

---

M. Urbano-Cuadrado  
Institute of Chemical Research of Catalonia ICIQ, Avinguda Països Catalans, 16, 43007 Tarragona, Spain

I. Luque Ruiz · M. Á. Gómez-Nieto (✉)  
Department of Computing and Numerical Analysis, University of Córdoba, Campus of Rabanales, Albert Einstein Building, 14071 Córdoba, Spain  
e-mail: mangel@uco.es



**Fig. 1** An example of graphs showing equal isomorphism

In chemistry, Quantitative Structure–Property Relationship (QSPR) and Quantitative Structure–Activity Relationship (QSAR) methodologies [1–3] try to correlate chemical structure differences and their respective changes of properties and biological activities. A high number of 2D and 3D descriptors accounting for topological, electronic or steric properties have been summarized in literature [4] as predictive variables.

Similarity measurements have been widely used in computational chemistry. Thus, similarity calculation algorithms support many approaches for both screening chemical databases and predicting physical–chemical properties [5]. In recent years, methods that correlate the 2D and 3D structural similarity between molecules—or between a molecular and a 3D grid probe—with their properties have been proposed based on the following chemical principle: “structurally similar molecules show similar properties and biological activities” [6–9].

Computational resources involved in the structural similarity calculation are great owing to the previous detection of (molecular) graph isomorphism. With the aim of overcoming this disadvantage, different methods for calculating graph isomorphism have been developed [10]. For this purpose, the use of binary fingerprints constitutes the basis of many similarity approaches. However, this similarity calculation based on the transformation of chemical structures into fingerprints has shown problems in QSPR models. Non-consideration of type, size and number of substructures produce low correlations between properties and fingerprint similarity values.

In addition to the computational cost commented, similarity measurements yield inconsistencies when different molecules show equal isomorphism, as can be seen in Fig. 1. When we calculate structural similarity between the A, B and C molecular graphs, the isomorphism consisting of 11 vertexes and 12 edges is equal for the three graphs. Similarity between any pair of the three molecular graphs is therefore equal ( $S_{A,B} = S_{A,C} = S_{B,C}$ ). Nevertheless, properties of the molecules represented by the A, B and C molecular graphs are different. This fact explains the low correlation achieved using structural similarity.

3D similarity calculations solve this problem by means of considering the different spatial conformations adopted by the molecules. Despite this advantage, some considerations should be taken into account. First, the geometrical optimization carried out is based on quantum or semi-empirical principles that surpass the role of topological calculation. Thus, calculations of electronic or steric fields of the molecules are often involved in the 3D similarity approaches. Therefore, these methods are more complex

and they require more computational resources than those related to topological calculations [11].

In order to overcome the above commented inconsistencies through a topological solution, we propose a new method for similarity/distance measurements between molecular graphs. This approach is supported on considering non-isomorphic subgraphs. Thus, if non-common molecular substructures are analyzed and measured with respect to a reference, a new structure-property space can be built and employed for predicting the properties of new compounds.

The development of a QSPR/QSAR model can be divided into three stages, namely: data obtaining, data analysis and model validation [12]. The first stage implies the selection and generation of molecular descriptors through different methods and software. The second stage includes the use of statistical applications for analyzing complex data arrays, being summarized in literature two main sorts of multivariate analysis methods: parametric approaches (multiple linear regression, principal components regression, etc.) and non-parametric approaches (artificial neural networks and genetic algorithms). The last stage consists of the use of several criteria and methodologies for validating the equations built in the previous stages. Several statistical parameters accounting for the correlation between predictors and properties, and the error involved in the predictive ability are employed in this step [13].

Taking into account the stages that constitute the QSPR development, this paper has been organized as follows: Sects. 2, 3 and 4 describes the data generation, the statistical technique and the validation process, respectively. The application of the proposed method to the prediction of sublimation enthalpies of Polychlorinated Biphenyls (PCBs) is described in Sect. 5. Finally, conclusions are highlighted in the last section.

## 2 Topological distances between non-isomorphic subgraphs

Similarity between two graphs  $G_A$  and  $G_B$  that represent the molecules  $A$  and  $B$  is expressed as follows:

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

where:  $S_{A,B}$  is a value within the range [0,1] that shows the similarity between the molecular graphs  $G_A$  and  $G_B$ ;  $I_{A,B}$  is the isomorphism between the graphs  $G_A$  and  $G_B$ ;  $A$  and  $B$  are the sizes of the graphs, and  $f$  is a function (algorithm) or approach that matches  $S$  and  $I$ . Thus, different similarity values can be obtained depending on the method employed for calculating the isomorphism between molecular graphs, namely: Maximum Common Edges Subgraph (MCES), Maximum Common Subgraph (MCS) or All Maximum Common Subgraphs (AMCS) [10]. When methods based on the transformation of graphs into fingerprints are used, different similarity values are also obtained, depending on the similarity index used and the characteristics of the fingerprint built [14].

As upon above stated, similarity measurements can lead to deviations in the correlation between molecular topologies and properties (QSPR). Our proposal takes into account the characteristics between subgraphs that do not form the isomorphism  $I_{A,B}$ .

We intend to search for relationships between variations of molecular properties and their differences according to structural topology.

Thus, we express the structural difference between two molecular graphs  $G_A$  and  $G_B$  as follows:

$$\Gamma_{A,B} = g(\text{td}[G_A - f(I_{A,B})], \text{td}[G_B - f(I_{A,B})]) \quad (2)$$

where:  $f(I_{A,B})$  has equal meaning to that shown in expression (1);  $G_A - f(I_{A,B})$  and  $G_B - f(I_{A,B})$  represent the subgraphs of  $G_A$  and  $G_B$ , respectively, that do not form the isomorphism  $I_{A,B}$ ;  $g()$  is a function aimed at obtaining a distance value between  $\text{td}[G_A - f(I_{A,B})]$  and  $\text{td}[G_B - f(I_{A,B})]$ ; and  $\Gamma_{A,B}$  is a metric technique that calculates the structural difference between the non-isomorphic subgraphs of  $G_A$  and  $G_B$ .

The method is therefore open with respect to several factors, namely: kind of isomorphism (functions  $f$ ), different descriptors or topological variables accounting for the non-isomorphic subgraphs in  $G_A$  and  $G_B$ , and the technique employed for measuring the distances between  $\text{td}()$  values. Thus, different QSPR models can be developed aimed at correlating  $\Gamma_{A,B}$  values with physical-chemical properties.

In this paper the model predictors have been obtained taking into account the following considerations:

- Isomorphism calculation (function  $f$ ) employed for this QSPR development was based on the MCS (Maximum Common Subgraph) approach.
- Topological descriptors ( $\text{td}$ ) were the Wiener, Hyper Wiener and Valence Overall Wiener (VOW) indexes [3,5]. The latter is supported by the Wiener calculation from the weighted distances matrix ( $D$ ) of a molecular graph. Elements  $D(i, j) = l$  of the distance matrix are replaced by elements  $D(i, j) = x$ , where  $x$  is the relative bond distance between the vertices (atoms)  $i$  and  $j$  with respect to the reference value corresponding to the C–C bond distance.
- Euclidean distance was the function  $g$  employed for measuring the difference between  $\text{td}[G_A - f(I_{A,B})]$  and  $\text{td}[G_B - f(I_{A,B})]$ .

A  $N \times N$  topological dissimilarity matrix ( $\Gamma$ ) can be built from the set consisting of  $N$  compounds. Each element  $\Gamma_{i,j}$  provides the topological distance between the non common subgraphs of the compounds  $i$  and  $j$  and it shows the same value as the element  $\Gamma_{j,i}$ . The higher differences there are between molecules, the nearer to 1 value for the  $\Gamma_{i,j}$  element. The diagonal of the matrix (elements  $\Gamma_{i,i}$ ) are equal to 0.

### 3 Data analysis with partial least squares regression (PLSR): reducing the distance space into latent variables

After data obtaining, the  $N \times N$  topological dissimilarity matrix ( $\Gamma$ ) can be considered as a data set formed by  $N$  training elements and  $N$  variables. Each variable accounts for the distances between a reference graph (represented by that variable) and the remaining objects. Thus, multivariate regression techniques have to be applied to the distance matrix in order to extract useful information. PLSR [15,16] was employed

due to several reasons. First, the original data space is transformed to a reduced system in which the fact of visualizing both trends and influences of the original variables on properties is a process more intuitive than that from the original space. Thus, the study of the number, type and characteristics of the PLS factors provides scientists with structured information of their systems.

Second, the fact of having a symmetrical matrix requires another technique different of multiple linear regression (MLR), which needs, algebraically, a system with more predictors than objects. Although several approaches attempt to solve this problem based on removing non significant variables, PLSR permits to work with modelling spaces consisting of more variables than objects.

And third, PLSR offers the advantages of considering variance of both the predictors and the properties for building the reduced space. This construction is more suitable to appropriate correlations between data and properties. Other techniques also based on the reduction of the variables only takes into account the predictor variance. For example, principal components regression (PCR) retains relevant factors that only explain the predictor set.

### 3.1 The PLSR modelling

Although several algorithms have been developed for computing partial squares components (the Non-linear Iterative Partial Least Squares NIPALS, the SIMPLS method, the PLS2 approach, etc., and their robust versions), the work methodology, described below, is common [17].

Be  $X$  and  $Y$  the matrixes that describe  $p$  observations and  $m$  properties, respectively, for  $n$  objects. A regression using factor extraction from data computes the factor score matrix  $T = XW$  for an appropriate weight matrix  $W$  (maximal  $X$  and  $Y$  data variance must be explained and overfitting must be avoided), and then considers the linear regression model  $Y = TQ + E$ , where  $Q$  is a matrix of regression coefficients (loadings) for  $T$ , and  $E$  is an error (noise) term.

Aimed at specifying  $T$ , two sets of weights  $w$  and  $q$  have to be found to create a linear combination of columns of  $X$  and  $Y$  such that their covariance is maximum. The goal is to obtain a first pair of vectors  $t = Xw$  and  $q = Yq$  with the constraints that  $w^T w = 1$ ,  $t^T t = 1$  (assures the orthonormality of the latent variables) and  $t^T q$  be maximal (reflects the maximal covariance structure between the predictor and property spaces). When the first latent vector is found, their contributions are subtracted from  $X$  and  $Y$  and this procedure is re-iterated until  $X$  becomes a null matrix. In this case, the number of latent variables is equal to the rank of  $X$ , and then, an exact decomposition of  $X$  and  $Y$  has been carried out.

Only a few latent variables are then considered for predicting the properties of new objects because of the overfitting phenomenon. Although the fact of using a high number of components means to correlate more accurately predictors and properties for the training set, this can also imply the error modelling. Thus, predictions carried out are affected by PLS factors that explain the error component of predictors. With the aim of both avoiding the error modelling and evaluating the predictive ability of the PLS models, the validation stage has a key role in the development of QSPR approaches.

#### 4 Validation of QSPR approaches: evaluating the predictive ability of the models

The model validation is the stage for both testing the predictive ability of the equations and comparing their efficiency with standards and criteria. Accuracy and precision, likely the two most important analytical properties in addition to representivity, are calculated. Since the QSPR development makes use of experimental data, the model will have maximum limits of accuracy and precision, which are given by the error involved in reference data obtaining. The optimal PLS factor number is another parameter also obtained in this stage aimed at removing noise modelling from the predictive equations.

The validation step can be studied with respect to three topics, namely: the methodology, the statistical parameters and the criteria employed for defining the quality of the models. Internal and external validations are the two methodologies employed in this stage. The former makes use of objects that have been yet employed for training and, contrary, external validation is based on the use of new objects. The kind of validation depends on the total number of compounds available for the QSPR development. Thus, when the set of objects is small, splitting this set into the training and test subsets involves the use of such a low number of objects that the predictive ability can not be modelled. In order to overcome this shortcoming,  $k$  fittings and the subsequent tests are carried out. For each cycle or fitting, the equations are built with the majority of the compounds and, then, are tested with the remaining objects. Several cycles are realized in such way that all the samples have been employed for testing. The final regression model is the average of the individual calibrations.

Full cross validation is a special kind of internal validation consisting of carrying out  $N$  cycles, where  $N$  is the number of objects, and  $N - 1$  and 1 objects compose the training and test sets for each cycle (also called Leave One Out, LOO).

The multiple determination coefficient ( $R^2$ ), the standard error in prediction (SEP), and the slope and bias (intercept) of the correlation analysis are the statistical parameters employed for evaluating the predictive ability. These parameters can be calculated by the following expressions (3–6), respectively:

$$R^2 = 1 - \frac{\sum (y_i - y'_i)^2}{\sum (y_i - \bar{y})^2} \quad (3)$$

$$\text{SEP} = \sqrt{\frac{\sum (y_i - y'_i)^2}{n - 1}} \quad (4)$$

$$\text{slope} = \frac{\sum (y_i - \bar{y})(y'_i - \bar{y}')}{\sum (y'_i - \bar{y}')^2} \quad (5)$$

$$\text{bias} = \bar{y} - \text{slope} \times \bar{y}' \quad (6)$$

The  $y_i$  and  $y'_i$  values are the experimental and predicted properties, respectively, and  $n$  is the number of compounds that compose the test set. In QSPR literature,  $R^2$  and

SEP are often called  $Q^2$  and SECV (Standard Error in Cross Validation), respectively, if they are referred to internal validation.

There are several criteria for determining the quality of chemometric approaches. Shenk et al. [18] proposed that  $R^2$  values higher than 0.90 indicate excellent precision, as well as SEP values lower than  $1.5 \times \text{EE}$  (Experimental Error).  $R^2$  values between 0.70–0.90 mean good precision, as do the SEP values among  $2\text{--}3 \times \text{EE}$ . On the other hand,  $R^2$  values lower than 0.70 indicate that the equation can only be used for screening purposes, which enable distinction between low, medium and high values for the measured parameter. If the  $R^2$  value is lower than 0.50, the equation only discriminates high and low values.

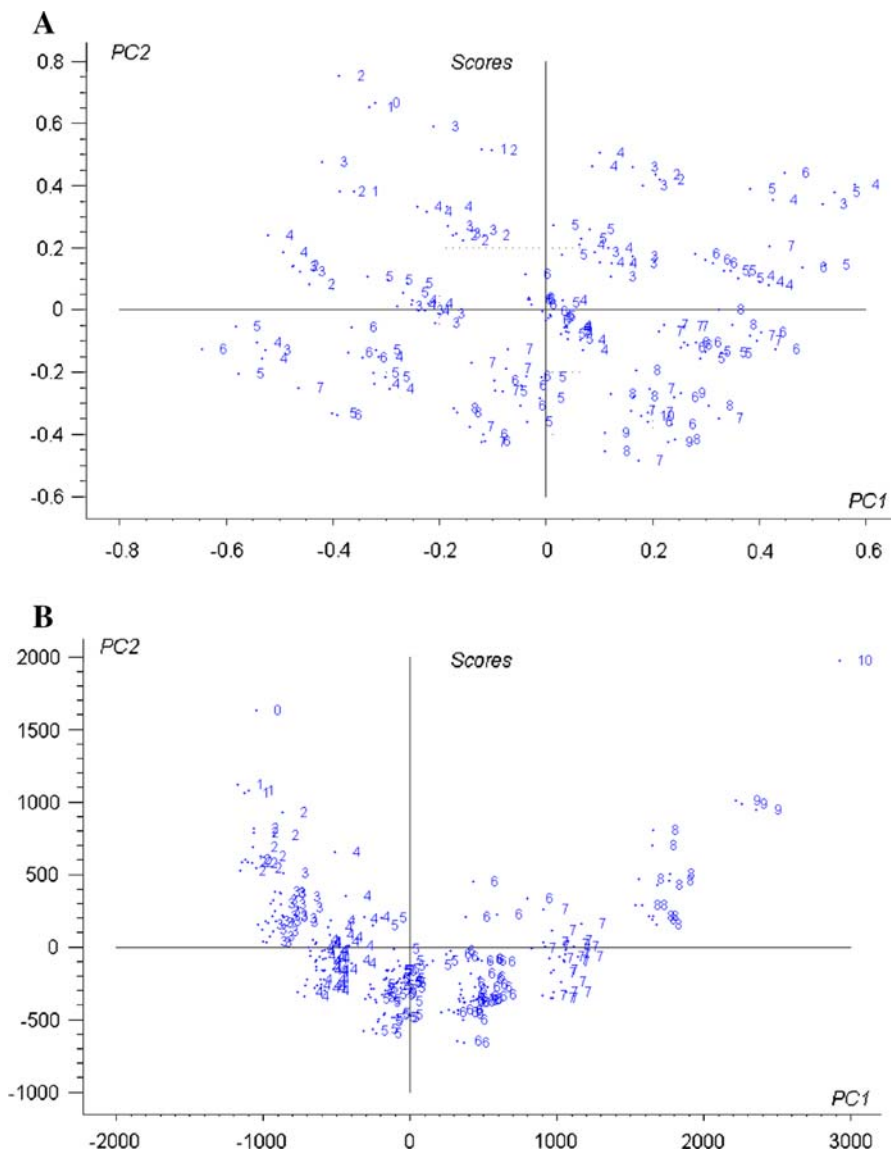
Taking into account the SEP value, it is also accepted by the scientific community that the limit for considering the equations as robust tools is  $1.5 \times \text{SEC}$  (Standard Error in Calibration). In addition, slope and bias values must be evaluated for testing if they are statistically equal to 1 and 0, respectively, at a given significance level (the most times at 0.5 or 0.25 % levels).

## 5 QSPR model for the prediction of sublimation enthalpy of polychlorinated biphenyls (PCBs)

PCBs have attracted the attention of the scientific community owing to the environmental problems related to organohalogen compounds. Non-flammability and chemical stability of these compounds, in addition to their lipophilicity, are responsible for their widespread problems. Several computational methods have been developed for estimating physicochemical properties—*n*-octanol/water partition coefficients [19], gas chromatographic retention times [20,21], relative heats of formation [22], lipophilicity, electron affinities and entropies [23], and sublimation enthalpy [24]—of the PCBs.

In this study, the method for calculating topological distances has been applied to a set of compounds consisted of 210 molecules—from biphenyl to decachlorobiphenyl, considering structural isomers for intermediate substituted biphenyls—. For this set of compounds, a structural dissimilarity matrix was built taking into account the non common subgraphs. Each element ( $i, j$ ) of this symmetrical matrix stores the  $\Gamma_{i,j}$  for each pair of elements. A similarity space ( $S$ ) was also obtained in order to compare our method with a similarity approach based on the MCES graph isomorphism calculation using the cosine index [25].

With the aim of carrying out an exploratory study dealing with the efficiency of the method we propose, Principal Components Analysis (PCA) was applied to the similarity ( $S$ ) and dissimilarity ( $\Gamma$ ) matrixes in order to make relevant hidden trends in data. Figure 2a and b show the score plots—the first two principal components—for the similarity and dissimilarity matrixes, respectively. The object identifier represents the number of the chlorines that substitute the biphenyl hydrogenous and varies therefore from 0 (the biphenyl) to 10 (the decachlorobiphenyl). A cluster for each class (each group of structural isomers is considered as a class and its objects show similar properties) was only formed when the dissimilarity space was employed. In addition, the first two components explain the 45 and 65% of the data variance (see the




















**Fig. 2** Score plots of the first two principal components for: **a** the similarity (S), **b** dissimilarity matrixes ( $\Gamma$ )

footnotes of the PCA plots) for similarity and dissimilarity values, respectively. These facts point out an improvement in the results considering non-common subgraphs and their differences.

The development of a QSPR model for predictions of sublimation enthalpy and the study of its efficiency is a first study to pre-evaluate potential approaches. The sublimation enthalpy  $\Delta_{\text{sub}}H_m(298.15 \text{ K})$  is a molecular property that provides information about the intermolecular forces that lead to the packing observed in the solid state.



**Table 1** Biphenyl and PCBs used in this study

Compound	Experimental	Predicted	Residual
	82.1	84.5	-2.4
	86.3	88.2	-1.9
	82.4	93.0	-10.6
	96.9	91.2	5.7
	105.1	95.0	10.1
	99.5	97.6	1.9
	103.6	99.9	3.7
	95.6	100.3	-4.7
	108.3	107.4	0.9
	101.0	100.6	0.4
	101.0	101.8	-0.8
	109.8	110.9	-1.1
	122.7	122.1	0.6
	101.5	104.4	-2.9
	122.7	127.0	-4.3
	114.2	111.6	2.6
	119.1	118.3	0.8

Experimental data have been obtained from the bibliography. Predicted results (full cross validation) and residuals obtained with the proposed model are also shown

Experimental values of this parameter for biphenyl and 16 PCBs were obtained from bibliography [6]. As Table 1 shows, these values were used for model training and testing.

As upon above commented, both the high correlation between predictors and the consideration of predictor and property variances justify the selection of PLSR for multivariate calibration. The low number of objects makes necessary to use internal validation of the predictive equation, and the statistical parameters employed were  $R^2$  (full cross validation), Standard Error in Cross Validation (SECV), slope and bias.

Univariate and multivariate analysis using topological descriptors (Wiener, Hyper Wiener and VOW indexes) and the similarity matrix of all the studied compounds,

**Table 2** Statistical results for the proposed method regarding with other studied approaches

Method	$R^2$ (full cross validation)	Standard error in cross validation (kJ/mol)	Slope	Bias (kJ/mol)
Univariate (TD)	0.56	8.62	0.81	19.42
Multivariate (Similarity)	0.72	6.42	0.92	7.90
CoMFA model	0.75	–	–	2.27
Multivariate ( $\Gamma$ )	0.87	4.51	0.98	1.79

respectively, have also been carried out in order to study the usefulness of considering both non-isomorphic subgraphs and Euclidean distance as the function  $g$ . Descriptor values employed in the univariate analysis were calculated taking into account the entire topology of the graph. The statistical parameters were also obtained after full cross validating the model. The VOW descriptor yielded the best univariate approach and this was then employed for multivariate analysis.

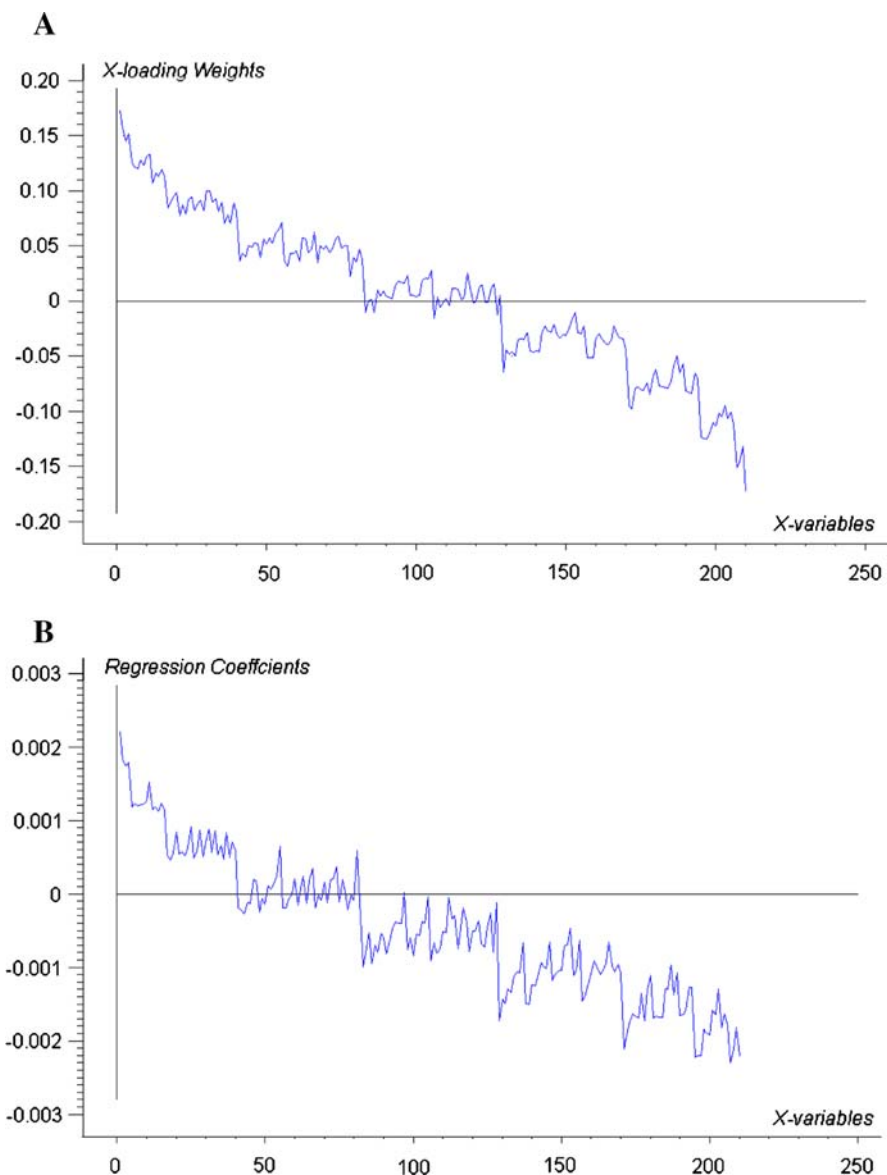
The fact of using a dissimilarity space lead to better results than those achieved when VOW values and the similarity matrix were considered. As Table 2 shows, PLSR applied to  $\Gamma$  matrixes increases  $R^2$  and slope values, and decreases SECV and bias numbers. Thus, accuracy and precision are improved when the method proposed is employed. According to chemometric criteria, the  $R^2$  values obtained for the similarity and dissimilarity matrixes permit the use of the model for quantitative predictive tasks, thus overcoming the screening role of the applications with  $R^2 < 0.70$ . Predicted values and residuals using the model proposed are shown in Table 1.

Figure. 3a and b show the loading weights for the first PLS factor and the regression coefficients of the original variables (distances), respectively. A general decrease of the weights and the coefficients with the size of the compounds can be observed in both plots. Thus, the higher differences between a molecule and the PCBs with few chlorine substitutions the higher value for the sublimation enthalpy. Although the trend is the decrease, there are a high number of consecutive variables showing low decreases and increases. Those variables accounts for the distances between structural isomers.

The method was also compared with the computational approach from which experimental values of sublimation enthalpy were obtained [24]. This approach is based on the *Comparative Molecular Field Analysis* (CoMFA) technique. Table 2 shows the values for the parameters used in our study and available from the above commented approach. These statistic values support the higher efficiency of our method.

## 6 Conclusions

In this paper new QSPR model based on both the consideration of graph isomorphism and the measurement of distances between the non-isomorphic subgraphs has been proposed.



**Fig. 3** Loading weights of the original variables (distances) for: **a** the first PLS factor, **b** the regression coefficients

Different kinds of isomorphism, topological invariants and distance approach can be considered in order to obtain a distance (dissimilarity) matrix for predicting properties. Thus, different models can be developed aimed at correlating  $\Gamma_{A,B}$  values with physical-chemical properties and/or biological activities.

A QSPR model for determining sublimation enthalpies of PCBs was built with the method here presented. The statistical values obtained in the PLSR model validation indicated an improvement on the predictive ability when the dissimilarity matrix was

employed as the predictor space. According to chemometric criteria, the QSPR model has the role of accurate and quantitative approach. The method showed better results than those obtained with complex 3D approaches. This fact would mean that advanced topological methods can correlate structures and packaging properties in an efficient way.

Besides this, distance measurements might be used to calculate “*fine similarity*” values between molecules. These corrected similarity values account not only for the structural similarity between two molecular graphs (subgraphs isomorphism) but also for the approximate similarity between the remaining non-isomorphic subgraphs. We are using distances and approximate similarity values for the development of new QSPR models and screening methods.

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

## References

1. H. Van de Waterbeemd (ed.), *Structure-Property Correlations in Drug Research* (Academic Press, Austin, 1996)
2. H. Kubinyi, *QSAR: Hansch Analysis and Related Approaches* (VCH, Weinheim, 1993)
3. C. Hansch, A quantitative approach to biochemical structure–activity relationships. *Acc. Chem. Res.* **2**, 232–239 (1969)
4. R. Todeschini, V. Consonni (ed.), *Handbook of Molecular Descriptors* (Wiley-VCH, Weinheim, 2000)
5. G.M. Downs, J.M. Barnard, in *Clustering and their uses in Computational Chemistry*, ed. by K.B. Lipkowitz, D.B. Boyd. *Reviews in Computational Chemistry*, vol 18. (Wiley-VCH, New York, 2003), pp. 1–39
6. O. Ivanciuc, A.T. Balaban, in *The Graph Description of Chemical Structures*, ed. by J. Devillers, A.T. Balaban. *Topological Indices and Related Descriptors in QSAR and QSPR* (Gordon and Breach Science Publishers, The Netherlands, 1999), pp. 59–167
7. D.H. Rouvray, A.T. Balaban, in *Chemical Applications of Graph Theory*, ed. by R.J. Wilson, L.W. Beineke. *Applications of Graph Theory* (Academic Press, New York, 1979), pp. 177–221
8. C.K. Hattotuwegama, A. Doytchinova, D.R. Flower, In silico prediction of peptide binding affinity to class I mouse major histocompatibility complexes: a comparative molecular similarity index analysis (CoMSIA) study. *J. Chem. Inf. Comput. Sci.* **45**, 1415–1423 (2005)
9. G. Klebe, U. Abraham, T. Mietzner, Molecular similarity indices in a comparative analysis (ComSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **37**, 4130–4146 (1994)
10. G. Cerruela García, I. Luque Ruiz, M.A. Gómez-Nieto, Step-by-step calculation of all maximum common substructures through a constraint satisfaction based algorithm. *J. Chem. Inf. Comput. Sci.* **44**, 30–41 (2004)
11. D. Robert, L. Amat, R. Carbo-Dorca, Quantum similarity QSAR: study of inhibitors binding to thrombin, trypsin, and factor Xa, including a comparison with CoMFA and CoMSIA methods. *Int. J. Quantum Chem.* **80**, 265–282 (2000)
12. A. Golbraikh, A. Tropsha, Be aware of  $Q^2$ ! *J. Mol. Graph. Model.* **20**, 269–276 (2002)
13. S. Wold, L. Eriksson, in *Statistical Validation of QSAR Results*, ed. by H. Van de Waterbeemd. *Chemo-metrics Methods in Molecular Design* (VCH, Weinheim, 1995), pp. 309–318
14. P. Willett, Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **38**, 983–996 (1998)
15. A. Hoskuldsson, PLS regression methods. *J. Chemom.* **2**, 211–228 (1988)
16. P. Geladi, B. Kowalski, Partial least square regression: a tutorial. *Anal. Chim. Acta.* **35**, 1–17 (1986)
17. S. Jong, SIMPLS: An alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.* **18**, 251–263 (1993)
18. J.S. Shenk, M.O. Westerhaus, *Calibration the ISI Way. In Near Infrared Spectroscopy: The Future Waves* (NIR Publications, Chichester, 1996), pp. 198–202

19. M. Makino, Prediction of *n*-Octanol/water partition coefficients of polychlorinated biphenyls by use of computer calculated molecular properties. *Chemosphere* **37**, 13–26 (1998)
20. M. Makino, Novel classification to predict relative gas chromatographic Retention Times and *n*-Octanol/water partition coefficients of polychlorinated biphenyls. *Chemosphere* **39**, 893–903 (1999)
21. M.N. Hasan, P.C. Jurs, Computer-assisted prediction of gas chromatographic retention times of polychlorinated biphenyls. *Anal. Chem.* **60**, 978–982 (1988)
22. J.A. Mulholland, A.F. Sarofim, G.C. Rutledge, Semiempirical molecular orbital estimation of the relative stability of polychlorinated biphenyl isomers produced by *o*-Dichlorobenzene pyrolysis. *J. Phys. Chem.* **97**, 6890–6896 (1993)
23. S.A. Kafafi, H.Y. Afeefy, H.A. Ali, H.K. Said, A.G. Kafafi, Binding of polychlorinated biphenyls to the aryl hydrocarbon receptor. *Environ. Health Perspect.* **101**, 422–428 (1993)
24. S. Swati Puri, J.S. Chickos, W.J. Welsh, Three-dimensional quantitative structure–property relationship (3D-QSPR) models for prediction of thermodynamic properties of polychlorinated biphenyls (PCBs): enthalpy of sublimation. *J. Chem. Inf. Comput. Sci.* **42**, 109–116 (2002)
25. X. Li, J. Lin, The valence overall wiener index for unsaturated hydrocarbons. *J. Chem. Inform. Comput. Sci.* **42**, 1358–1362 (2002)