Effects of Molecular Representation in Predicting the Biological Activity using SVM and PLS Approaches

Gonzalo Cerruela García, Irene Luque Ruiz, Nicolás García Pedrajas, Miguel Ángel Gomez-Nieto

University of Córdoba. Department of Computing and Numerical Analysis. Campus de Rabanales. Albert Einstein Building. E-14071 Córdoba, Spain, {gcerruela, ma1lurui, npedrajas, mangel}@uco.es

Abstract

In this work we study and analyze the behavior of different representational spaces for molecular activity prediction. Representational spaces based on fingerprint similarity, structural similarity using maximum common subgraphs (MCS) and all maximum common subgraphs (AMCS) approaches are compared against representational spaces based on structural fragments and non-isomorphic fragments (NIF), built using different molecular descriptors. Support vector machine is used to study the influence of molecular representation in the dataset classification and PLS regression is proposed to construct a QSAR model for the molecular activity prediction.

Keywords: Molecular activity prediction, QSAR models, SVM, PLS regression.

1 INTRODUCCTION

In the pharmaceutical industry one of the most important steps for the molecular activity prediction, and in the construction of quantitative structure-property relationships (QSPR) and quantitative structure-activity relationship (QSAR) models is the data set representation. An adequate modeling of the structural and physicochemical features of chemical compounds is crucial for the prediction of biological molecular activity and to understand the molecular mechanism of a particular biological activity from the derived statistical models [1]. Different spaces have been used to represent the molecule data set, based on molecular descriptors [2] or considering structural information of the molecules, such as methods based on structural and fingerprint similarity [3], as well as structural fragments [4] extracted during the fragmentation process of the entire data set.

In the last decade, several graph representations have been introduced to depict theoretical biological data. Thus, complex networks based on graph theory are used to represent the structure and dynamics of different large biological systems such as protein--protein interaction networks. Complex networks are made up of nodes and edges usually representing the similarity and dissimilarity relationships between the nodes and placed in the space with or without geometrical constraints. For instance, in these networks, aminoacids could play the role of nodes and edges express spatial contact between two aminoacids or nucleotides could play the role of nodes and the edges are sequence neighbors or represent hydrogen bond thus denoting the secondary structure of RNAs. Such complex networks are used to compute various descriptors to describe the structure of drugs, proteins, or large bio-systems [5].

Other approaches are based on the similarity between the molecule dataset. Similarity measurements [6] are also obtained from graph representation of the molecules through an algorithm to calculate the “resemblance” between two given molecular graphs. These algorithms, named matching or isomorphism algorithms, extract the commonality between the molecular graphs representing the molecules. Subsequently, counts of these commonalities are used for the calculation of various similarity measures.

The calculation of the common subgraphs between two given molecular graphs has a high computational cost, and different solutions have been proposed to extract different measures of subgraph isomorphism, and the maximum common subgraph (MCS) is the most used. The MCS between two given molecular graphs is the maximum (largest number of nodes and edges) connected subgraph common to the two matched graphs. Another usual measure of isomorphism is the maximum common edges subgraph (MCES) or the maximum overlapping set (MOS) which is the maximum (not necessarily connected) subgraphs or cliques (more than one) common to the two matched graphs [7].

The array representations of molecule structure are named fingerprints [8] and they have a more efficient algorithm behavior [9]. Fingerprints are binary arrays that are generally built from the information stored in the molecular graph. Different fingerprint models have been proposed based on the extraction of structural elements of the molecular graphs and the application of different hashing algorithms to determine the bits set to 1 in the fingerprint. Although in the process of the building of the fingerprint the entire information stored in the molecular graph cannot be translated, fingerprint representations have shown huge efficiency in diverse fields of computational chemistry.

In this work we have used a SVM algorithm with feature selection method (RFE) [10] for the analysis of the behavior of the different representational spaces in the prediction of molecular biologic activity. Feature selection methods have been introduced for improving the classification performance of statistical learning methods and for selecting meaningful features when discriminating two data sets. The theory and applications of support vector machine in chemistry have been extensively reviewed [11]. SVM is used as the statistical learning method to study the influence of molecular representations in the activity classification, and PLS regression is used to construct a QSAR model for the molecular activity prediction.

The paper is organized as follows: in section 2, we describe the theoretical basis of the isomorphism algorithms, and similarity concepts, we present the techniques used for the generation of the different representational spaces based on structural information of the molecule dataset. In section 3 we analyze the experimental results and the behavior of different representational spaces in molecular classification and to construct an activity prediction model using PLS regression. Finally, the main conclusions are summarized.

2 THEORETICAL FOUNDATIONS

2.1. DESCRIPTOR-BASE REPRESENTATION

Molecule data sets can be represented as an array of variables or descriptors [2] representing some structural property of the chemical compound in order to extract a structural similarity measurement. Thousand of molecular descriptors have been proposed in the literature, most of them related to some structural or topological property corresponding to the molecular graph representing the molecule. For the calculation of descriptor-based similarity, molecules are represented as equation (1) and the two arrays are compared using mathematical or statistical procedures. However different hints are necessary to be considered:

$$M_A = \left(d_1^A, d_2^A, d_3^A, \cdots, d_n^A\right)$$
$$M_B = \left(d_1^B, d_2^B, d_3^B, \cdots, d_n^B\right) \quad (1)$$

The number and order of the descriptors has to be the same in the arrays M_A and M_B. The magnitude of the descriptors has to be normalized.

It is mandatory to eliminate the relationships between descriptors in order to diminish the number of variables, being necessary to perform a previous statistical analysis.

2.2. FINGERPRINT REPRESENTATION

In this case the compounds are represented as vectors of 0 and 1, indicating the presence or absence in the molecule of a particular substructure (see Figure 1).

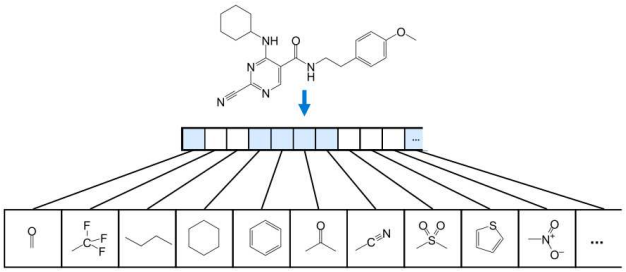

Figure 1. Molecular Fingerprint extraction

In fingerprint generation process the following parameters have been adjusted [12]:

- ✓ *Fingerprint length*, establishing the length of the binary array representing the fingerprint. This parameter takes values between 64 and 1024 bits, a value of 512 being the lowest value recommended for a better characterization of the molecular structure.
- ✓ *Bits to be set for patterns*, after detecting a pattern, some bits of the bit string are set to 1. The number of bits used to code patterns is constant.
- ✓ *Path length*, determining the high value of the size of molecular fragments extracted from the molecules in the generation process of the fingerprint.

By increasing the Fingerprint Length, the capacity for storing information about the molecules is greater. In addition, it decreases fingerprint darkness (a dark fingerprint has a high number of 1), and therefore also the probability of bit collisions also, which is beneficial. An excessive increase in Fingerprint Length may raise the necessary disk space, the size of the structure cache and the processing time.

By increasing the number of bits to be set for patterns, the coded information derived from a pattern increases. The fingerprint darkness rises, and as long as the bit collision number is not too high, the stored information increases, which is beneficial for the efficiency of the screening.

2.3. MOLECULAR FRAGMENTATION

This approach have been proposed in order to consider the fragments information of the molecules' dataset in the construction of prediction models. In this process, fragments made of different paths of molecule structures are extracted by applying different heuristics (e.g. minimum and maximum path size, considering hybridization or not, maintaining the cycle structure or not, etc.). In this way, the dataset can be represented again by a $Z=M \times F$ matrix, with M the dataset cardinality, and

F the number of the different generated fragments. Now, again using any molecular descriptor we can fill this matrix, storing in each matrix element *(i, j)* the descriptor value of the fragment *j*, for the molecule *i*.

2.4. MOLECULAR GRAPH ISOMORPHISM

Graph isomorphism has been widely used in computational chemistry. These models consider that structurally similar molecules should show similar properties and biological activities.

Thus, the molecular structure is represented through a graph called molecular graph. A molecular graph G is a connected and non-directed graph, where the nodes represent the atoms and the edges represent the bonds in the molecule. Determining whether two graphs G_A and G_B are the same (graph isomorphism) is a difficult computational task. Also difficult is the subgraph isomorphism problem, in which one determines which subgraph of G_A is a subgraph of G_B or vice versa.

The detection of subgraph isomorphism, and especially the detection of the maximum common substructure (MCS), has been applied to a great variety of fields, such as similarity-based problems, design and synthesis of products, clustering, QSPR/QSAR, etc.

A graph G is called a maximum common subgraph (MCS) of two graphs G_A and G_B , if there is no other common subgraph of G_A and G_B that has more nodes than G (node induced subgraph). According to this definition, MCS is not necessarily unique for two given graphs and it is always a connected graph representing the maximum clique between G_A and G_B . Another definition of subgraph isomorphism considers maximum overlapping set, MOS) which contains the maximum number of common edges between two given molecular graphs G_A and G_B.

For the last years, several algorithms have been proposed for the calculation of the MCS trying to reduce its computational cost. However, while all approaches calculate the MCS and MCES, our proposal allows to calculate the AMCS (All Maximum Common Subgraph) [7] . Because of MCS can be not unique for two given graphs, this algorithm has several steps: obtaining the MCS from the higher to the lower size, generating a set of connected MCS subgraphs related each order in a isomorphic AMCS and a non-connected common subgraph. Thus, the AMCS is a set of maximum and common subgraphs to G_A and G_B where: a) all the graphs of the AMCS set are connected graphs, b) the elements of the AMCS set are MCS^k subset that maximize the next MCS^{k+1}, more details about AMCS can be found in [7].

2.5. CHEMICAL SIMILARITY

Similarity is a fuzzy concept that attempts to measure the "resemblance" between two real or abstract objects. For more than fifty years researchers have applied this concept to different problems in chemistry. Database clustering and screening and the development of predictive models of physicochemical properties and biological activities of substances are clear examples of the application of similarity measures.

The different approaches to define chemistry similarity are based on: a) the review of structural commonalities or differences directly from the structural information of the chemical compounds, usually based on the topological representation of the structure, and b) the parameterization of the chemical structure by means of several specific property values and with the assistance of some mathematical analysis (statistical, geometrical, etc.) or artificial intelligence methods (pattern recogntion, neural network, etc.) in order to establish a numerical model. For the former, molecules are represented by molecular graphs that define the isomorphism relationships; then, a similarity is calculated based on the number of common nodes and edges of the common subgraph (MCS, MOS) regarding the total number of these elements in the matched graphs. Several similarity indices have been proposed in the chemistry literature, such as the Tanimoto, Ochiai, Sorensen, and Kulczynski indices [6]. For the second, molecules are represented by an array of properties or variables, called "descriptors", so that a molecule can be considered a point in a multidimensional "descriptor space", then two molecules, represented by their corresponding descriptors arrays are compared through a distance measure (e.g. Euclidean, Mahalanobis) between both descriptors spaces. Variables or descriptors can be extracted from the whole molecule or frequently using the fingerprint representation of the molecules.

2.6. HIERARCHICAL REPRESENTATION

This representation was based on the MCS calculation of all the pair wise molecules of the dataset in order to generate a hierarchical representation and to classify the dataset according with the different kernel substructures. The MCS built, shown in Figure 2, has the following characteristics: a) the root node contains the MCS common to all the dataset, b) each node also contains a MCS substructure common to all the child nodes, and c) finally, the leave nodes contain the molecules of the dataset.

The information contained in the nodes and edges of the MCST is used for building a weighted maximum common subgraph tree (WMCST), based on molecular descriptors [13]. The calculation of the molecular descriptor of nodes and edges allows the building of a weighted representation of the MCST. As the edges of the MCST store the NIF fragment from the parent and child nodes, the descriptor value of the NIF fragments can be used as a distance measure, that is, the cost of getting to the child node from the parent node. Thus, the WMCST can be represented by a $Y=N \times I$ matrix, being N the number of

WMCST nodes and *I* the number of different NIF substructures.

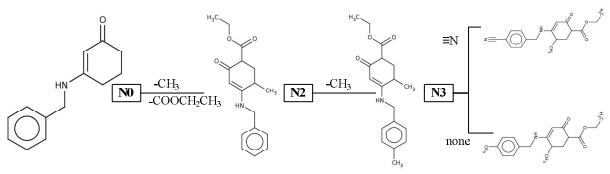

Figure2. Hierarchical representation.

The diagonal elements of this matrix store the weight of the nodes, and the rest of the elements store the weight of NIF fragment for each edge of the tree. Note that, while in the Z matrix all *i* elements for a same *j* element store the same value, in the Y matrix, the descriptor value for a NIF(j) can be different for two different WMCST nodes (i).

3 RESULTS AND DISCUSSION

In this work we have selected a dataset corresponding with 1,4-benzoquinone derivatives [14]. This family of substances has been studied due to its antifungal activity in humans. In recent years, the study of antifungal agents has attracted considerable interest due to an increase of mycotic infections and frequent accounts of tolerance to the known treatment. Recent studies have found that the presence of 1,4-benzoquinone derivatives in the structure of antifungal agent UHDBT 1 (5-n-undecyl-6-hydroxy-4,7-dioxobenzothiazole), one of the most effective antifungal agents known, improves considerably its activity. The experimental results for the representations, only based on the calculation of descriptors were not good, therefore have not been included in the tables of this section.

3.1. CLASSIFICATION USING SUPPORT VECTOR MACHINE ALGORITHM

In this process two SVM kernels have been tested: linear and Gaussian. Because SVM is very sensitive to the learning parameters; we have performed a 10-fold cross-validation for obtaining their values. Each time an SVM was applied, we tried a linear kernel with C in {0.1, 1, 10}, and a Gaussian kernel with C in {0.1, 1, 10} and gamma in {0.0001, 0.001, 0.01, 0.1, 1, 10}, testing all the 21 possible combinations. The best set of parameters obtained by the cross-validation process was then applied to learn the classifier.

In the study of the classification of the dataset using the different representational spaces, we have considered the following statistics for the analysis of the results:

$$Accuracy\ (A) = a + c / N \quad (2)$$

$$Sensitivity\ (S) = a / (a + d) \quad (3)$$

$$Specificity\ (s) = a / (c + b) \quad (4)$$

$$APV = a / (a + b) \quad (5)$$

$$IPV = c / (c + d) \quad (6)$$

where: *N* is the number of samples, *a*: is the number of active samples classified correctly, *b* is the number of active samples classified incorrectly, *c* is the number of inactive samples classified correctly, *d* is the number of inactive samples classified incorrectly, *APV* is the predictive value for active samples, *IPV* is the predictive value for inactive samples.

As observed in Table 1 the best results for the tested methods based on the use of classical similarity matrixes are obtained when the AMCS-based similarity is used. Tanimoto similarity index has been used for all the methods tested. Fingerprint-based similarity shows the worst results, although better than that from the literature [14]. In this case, compound 33 is classified as false inactive and four compounds (57, 65, 67, 71) as false active.

Compound 33 has the fluorine substituent in the para-position in the R2 (aniline) substituent. However, similar compounds like 32 or 34 are inactive compounds, the algorithm has a fault in the prediction. Regarding the errors in the classification on inactive compounds, we have noticed that compound 71 is the only compound that includes a double naphthalene substituent.

Compounds 65 and 67 are similar to compound 66, however they belong to a different class because of, again, the position of the fluorine substituent. The error in classifying the compound 57 maybe due to the low value of the pMIC 50 because it has not any substituent at the benzyl ring.

When MCS-based similarity matrix is used, compound 33 is again classified as false inactive, but in this case the compounds 28 and 61 are classified as false active.

Table 1. Results of SVM classification for different representation spaces (*Not external-Validation Used*)

Model	Active (54)		Inactive (20)		A	S	s	APV	IPV
	C	I	C	I					
Fingerprint	53	(33)	16	(57,65,67,71)	0.93	0.93	0.94	0.98	0.80
MCS	53	(33)	18	(28,61)	0.96	0.96	0.95	0.98	0.90
AMCS	54	None	20	None	1.00	1.0	1.0	1.00	1.00
Fragment (Z matrix)	54	None	20	None	1.00	1.0	1.0	1.00	1.00
NIF (Y matrix)	54	None	20	None	1.00	1.0	1.0	1.00	1.00

Compound 28 presents a low activty regarding with other similar compounds as 27, generating an erroneous classification. Regarding compound 61, a similar behavior to the aformentioned for the fingerprint matrix occurs. The ortho- position of the chlorine substituent provides a low activity, however, if the position is para- or meta- as in compound 62 the compounds are active.

Thus, we have observed that when fingerprint or MCS-based matrixes are used, some classification errors are found due to the existence of quite similar compounds that present a high difference of the experimental property. However, as it is observed in Table 1, when the AMCS-based similarity matrix is used, the algorithm is able to classify properly the entire dataset. The AMCS method provides more information than the MCS, generating finer similarity measures.

When fragments-based method is used, a complete classification of the dataset is obtained (see Table 2). The results are independent of the molecular descriptor used for building the Z matrix, because the presence or absence of specific fragments determines the class in which the molecule would be classified.

3.2. ACTIVITY VALUE PREDICTION USING PARTIAL LEAST-SQUARE REGRESSION.

Partial least-square regression (PLSR) [15] is one of the traditional statistical methods used for the development of QSAR models, so PLSR can be used with representational spaces (such as similarity matrices) in which the number of objects (molecules) and variables is the same. Moreover PLSR analysis reduces the data representational space, considering the variance of both predictors and properties.

Hence, the analysis of the PLS factors obtained allow us to perform the analysis of the multivariate system considering the trends and influences of the original variables on properties.

For the development of QSAR model we used the leave-one-out(LOO) cross-validation technique, evaluating as the result, the determination coefficient Q^2 (cross-validated R^2) and the standard error (SECV). Values of Q^2 greater than 0.50 and values of SECV lower than the standard deviation of the dataset predicted property would be the accepted threshold for proposing a model with a statistical meaning.

In order to analyze the behavior of the different representational spaces studied, the dataset was divided in two groups. The training group was composed by 55 molecules, representing 75 % of the entire dataset and the remaining 19 molecules, representing 25 % of the entire dataset were selected for the validation stage of the prediction model. Table 2 shows the results obtained.

As seen in Table 2, the similarity-based representational spaces (fingerprints, MCS and AMCS) have Q^2 values higher than 0.70. It is accepted within the QSAR community that: a) $Q^2 > 0.90$ indicates an excellent precision, b) if Q^2 values are between 0.70 and 0.90, that would mean that the model has a good precision, c) $Q^2 < 0.70$ indicates that the equation can only be used for screening purposes, which enable distinction between low, medium, and high values for the measured parameter, and d) if $Q^2 < 0.50$, the model only discriminates between high and low property values.

Best values of slope, bias and Q^2 were observed for fingerprint-based models in the training stage. However, the fingerprint-based model founds an outlier (molecule 24) and it needs more factors than MCS or AMCS-based models.

Also, for all models, the SECV is lower than the standard deviation of the dataset. However, the prediction capacity of the model is best for an AMCS based representational space. This model does not generate outliers and R^2, slope, bias and SEP values are close to the ideal.

Table 2. PLS QSAR prediction model results

Representation	Training					Test				
	Slope	Bias	Q^2	SecV	Outliers	Factors	Slope	Bias	R^2	Sep
Similarity Based										
Fingerprint	0.97	0.08	0.82	0.15	24	10	0.89	0.48	0.76	0.22
MCS	0.92	0.35	0.72	0.20	None	10	0.85	0.60	0.82	0.20
AMCS	0.92	0.33	0.71	0.20	None	6	0.98	0.06	0.82	0.18
Fragments Based										
Hyper Wiener	0.99	0.06	0.52	0.23	24,25,26	10	1.07	-0.35	0.87	0.16
Kier and Hall	0.86	0.59	0.69	0.18	22, 24,25,26	10	0.84	0.69	0.81	0.20
Randic (X^1)	0.86	0.60	0.69	0.18	22, 24,25,26	10	0.84	0.70	0.81	0.20
Schultz	0.95	0.22	0.63	0.20	24,25,26	10	0.97	0.15	0.83	0.18
Szeged	0.94	0.27	0.66	0.19	22, 24,25,26	10	1.03	-0.09	0.84	0.18
Wiener	0.95	0.24	0.62	0.20	24,25,26	10	0.99	0.03	0.83	0.18
NIF Based										
Hyper Wiener	0.94	0.26	0.85	0.13	23, 24, 26, 44, 54	9	0.96	0.17	0.87	0.15
Kier and Hall	0.95	0.20	0.68	0.18	22, 23, 24, 25, 26	3	0.98	0.01	0.78	0.21
Randic (X^1)	0.92	0.32	0.78	0.16	24, 26, 44	10	0.93	0.27	0.86	0.16
Schultz	1.13	-0.63	0.72	0.17	22, 23, 24, 25, 26	10	1.43	-2.06	0.72	0.26
Szeged	1.14	-0.68	0.76	0.16	22, 23, 24, 25, 26, 44	10	1.42	-2.01	0.72	0.27
Wiener	0.95	0.20	0.84	0.14	23, 24, 26, 44, 53	10	0.97	0.10	0.88	0.15

Results using fragments-based models as the representational space of the dataset are also shown in Table 2, for those molecular descriptors that show the best predictions. If we compare them versus similarity-based models, we observe lower statistics for the fragments-based models in the training stage: Q^2 values are lower (between 0.52 and 0.66) and the number of outliers, factors and SECV are higher. Fragments-based models require 10 factors for building the equation and 3-4 molecules are eliminated for the model. Again, molecule 24 is determined as an outlier as well as molecules 22, 25 and 26. These molecules are those presenting a simple substituent in the para-position in the *R2* radical (aniline).

Based on the results from Table 2, we conclude that NIF-based models improve considerably the results obtained in the previous analysis. The training models are better for all descriptors considered. Although the number of factors is similar for some descriptors, for the Kier and Hall descriptor only three factors are needed. Also, Q^2 values are always higher than 0.7, and for the Wiener-based descriptors this statistical is higher than 0.8. The number of outliers is slightly higher (3-5). Again, molecules 22, 24, 25 and 26 are defined as outliers, and also molecules 23, 44, 53 and 54 for some descriptors.

4 CONCLUSIONS

Molecular graph is a widely used mathematical mechanism for the representation of structural information of chemical compounds. By applying graph theory it is possible to extract structural information from molecules represented by means of a molecular graph. Thus, molecular descriptors can be obtained from molecular graphs, and then their values can be used in classification and prediction of molecular properties and biological activities of chemical compounds.

Moreover, molecular graphs can be used to obtain similarity measures between molecules, representing the resemblance between their structural characteristics, for a later use of these similarity measures in QSAR and classification activities.

Thus, structural based approaches using molecular descriptors and structural similarity have been widely used in computational chemistry during the last fifty years. In this paper we present and compare models previously proposed by the authors with other classical models of molecular representation. Descriptor, similarity and fragments based approaches have been studied and compared, and we presented the advantages and inconveniences of each of them.

References

[1] Irene Luque Ruiz, Gonzalo Cerruela García, M.Á. Gómez-Nieto, Structural Similarity based Approaches for the Development of Clustering and QSPR/QSAR Models in Chemical in: Matthias Dehmer, Kurt Varmuza, Danail Bonchev, F. Emmert-Streib (Eds.) Statistical Modelling of Molecular Descriptors in QSAR/QSPR, Wiley-VCH Verlag GmbH & Co. KGaA, 20012.

[2] M. Thomsen, L. Carlsen, Evaluation of empirical versus non-empirical descriptors, SAR QSAR Environ. Res., 13 (2002) 525-540.

[3] R. Van Deursen, L. Blum, J.-L. Reymond, Visualisation of the chemical space of fragments, lead-like and drug-like molecules in PubChem, J. Comput. Aid. Mol. Des., 25 (2011) 649-662.

[4] K.Z. Myint, C. Ma, L. Wang, X.Q. Xie, Fragment-similarity-based QSAR (FS-QSAR) algorithm for ligand biological activity predictions, SAR QSAR Environ. Res., 22 (2011) 385-410.

[5] R. Concu, M.A. Dea-Ayuela, L.G. Perez-Montoto, F.J. Prado-Prado, E. Uriarte, F. Bolás-Fernández, G. Podda, A. Pazos, C.R. Munteanu, F.M. Ubeira, H. González-Díaz, 3D entropy and moments prediction of enzyme classes and experimental-theoretic study of peptide fingerprints in Leishmania parasites, Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 1794 (2009) 1784-1794.

[6] N. Nikolova, J. Jaworska, Approaches to Measure Chemical Similarity – a Review, QSAR & Combinatorial Science, 22 (2003) 1006-1026.

[7] G. Cerruela García, I. Luque Ruiz, M.Á. Gómez-Nieto, Step-by-Step Calculation of All Maximum Common Substructures through a Constraint Satisfaction Based Algorithm, Journal of Chemical Information and Computer Sciences, 44 (2003) 30-41.

[8] P. Guttiérrez Toscano, F.H.C. Marriott, Unsupervised classification of chemical compounds, J. R. Statist. Soc. C, 48 (1999) 153-163.

[9] M. Vogt, J. Bajorath, Bayesian screening for active compounds in high-dimensional chemical spaces combining property descriptors and molecular fingerprints, Chem. Biol. Drug Des., 71 (2008) 8-14.

[10] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning, 46 (2002) 389-422.

[11] O. Ivanciuc, Applications of Support Vector Machines in Chemistry, in: Reviews in Computational Chemistry, John Wiley & Sons, Inc., 2007, pp. 291-400.

[12] JChem 6.1.4 , 2013, ChemAxon (http://www.chemaxon.com).

[13] G. Cerruela García, I. Luque Ruiz, M.A.n. Gómez-Nieto, Analysis and Study of Molecule Data Sets Using Snowflake Diagrams of Weighted Maximum Common Subgraph Trees, Journal of Chemical Information and Modeling, 51 (2011) 1216-1232.

[14] S.-Y. Choi, J.H. Shin, C.K. Ryu, K.-Y. Nam, K.T. No, H.-Y. Park Choo, The development of 3D-QSAR study and recursive partitioning of heterocyclic quinone derivatives with antifungal activity, Bioorganic & Medicinal Chemistry, 14 (2006) 1608-1617.

[15] D.A. Konovalov, L.E. Llewellyn, Y. Vander Heyden, D. Coomans, Robust Cross-Validation of Linear Regression QSAR Models, Journal of Chemical Information and Modeling, 48 (2008) 2081-2094.