DOI: 10.1002/minf.201100098

Navigating Drug-Like Chemical Space of Anticancer Molecules Using Genetic Algorithms and Counterpropagation Artificial Neural Networks

Mehdi Jalali-Heravi*[a] and Ahmad Mani-Varnosfaderani[a]

Abstract: A total of 6289 drug-like anticancer molecules were collected from Binding database and were analyzed by using the classification techniques. The collected molecules were encoded to a diverse set of descriptors, spanning different physical and chemical properties of the molecules. A combination of genetic algorithms and counterpropagation artificial neural networks was used for navigating the generated drug-like chemical space and selecting the most relevant molecular descriptors. The proposed method was used for the classification of the molecules according to their therapeutic targets and activities. The selected molecular descriptors in this work define discrete areas in chemical space, which are mainly occupied by particular classes of anticancer molecules. The obtained structure-activity relationship (SAR) patterns and classification rules contain valuable information, which help to screen the large databases of compounds, more precisely. Such rules and patterns can be considered as virtual filters for mining the large databases of compounds and are useful in finding new anticancer candidates.

Keywords: Cancer chemotherapy · General structure-activity relationships · Classification · Virtual screening · Drug discovery

1 Introduction

Cancer is the major public health problem in the United States of America (USA) and many other parts of the world.[1] Over one million new cases of cancer patients and more than 0.5 million cancer related deaths have been reported in the USA in 2010.[2] There are over 100 different types of cancer, and each is classified by the type of the cell that is initially affected. However, the common denominator for different types of cancer is unrestricted proliferation of tumor cells.[3] Various kinds of treatments, such as surgery, radiotherapy, proton therapy, gene therapy and chemotherapy have been proposed for controlling the abnormal proliferation of the cancerous cells.[4] Chemotherapy has been found to be an effective way for controlling tumor propagation in the body and is extensively used for the treatment of the cancer. During the chemotherapy, many combinations of medicines are used to weaken or destroy the cancerous cells.

Design and synthesis of new anticancer molecules with potent activity and less side effects has received great attention in recent years in the field of cancer chemotherapy.[5] Development of new anticancer drugs is commercially important and various pharmaceutical companies[6] and industries are focused on discovering new and more effective small molecules for the treatment of the cancer. Every year, a great number of articles about the biological synthesis and evaluation of new anticancer molecules are published. These studies tend to optimize the structure of the molecules for a better activity, with the aid of the principles of the structure-activity relationship (SAR) hypothesis.[7] To identify new chemical entities, for more effective treatment of the cancer, medicinal chemists may systematically search in the chemical space.[8] As the number of the possible drug-like molecules in existing chemical space (the total descriptor space that encompasses all of the small carbon-based molecules that could, in principle, be created) is rather infinite, the process of compound selection and priorization has become crucial in drug discovery. The techniques of ligand based virtual screening (LBVS) are among popular methods for selecting available compounds from large databases. These methods are very helpful for performing systematic search in existing chemical space and finding appropriate sub-spaces for exploration. The LVBS techniques are based on the concept of ligand similarity and generally are divided to molecular similarity and compound classification techniques.[9] Compound classification techniques use clustering algorithms, dimension reduction methods and partitioning techniques, such as neural networks, support vector machines and kernel based approaches for classifying the molecules into different categories. The classification and clustering techniques are very

[a] *M. Jalali-Heravi, A. Mani-Varnosfaderani*
Department of Chemistry, Sharif University of Technology
P.O. Box 11155-9516, Tehran, Iran
tel: +98-21-66165315; fax: +98-21-66012983
**e-mail: jalali@sharif.edu*

Supporting Information for this article is available on the WWW under http://dx.doi.org/10.1002/minf.201100098.

popular tools for exploring the chemical space of large and diverse datasets.[10] These techniques are very useful for the characterization of discrete areas of the chemical space that are occupied by compounds with considerable affinity for specific biological molecules. The classification models can help for screening large compound databases and can replace a considerable amount of laborious work in drug discovery projects by a more unbiased approach.

In recent years, the number of synthesized and evaluated small molecules, as potential anticancer agents, has been grown very fast. Statistical analysis of these compounds can help for deriving general SAR patterns and designing new anticancer drugs. The main aim of the present work was collecting different anticancer molecules and analyzing them by using the classification techniques. The classification models characterize different clusters of anticancer molecules in the space of the selected molecular descriptors. Characterization of such areas results in some simple rules, which accelerate the screening of large databases of compounds. Generally, we were interested to find specific patterns of molecular descriptors, which are unique for a particular class of anticancer agents. In this respect, a total of 6289 drug-like anticancer molecules was collected from the Binding database[11] and was subjected to SAR monitoring procedure using classification techniques.

The collected molecules in this work consisted of histone deacetylase inhibitors (HD-INs),[12] aromatase inhibitors (AR-INs),[13] endothelical growth factor receptor inhibitors (EGFR-INs)[14] and matrix metallo-proteinase 1 inhibitors (MMP-INs).[15] EGF tyrosine kinase receptors and HD, MMP-1 and AR enzymes are highly distributed on tumors and their inhibitors are proposed as potent anticancer agents with considerable selectivity over cancerous cells. These inhibitors show less destructive side effects compared to the conventional chemotherapy medicines.[5]

The method of counterpropagation artificial neural network (CPANN)[16] was used for the classification of compounds. CPANN is the supervised format of self organized maps (SOM)[17] and has great ability for clustering similar objects in a dataset. This method is proven to be helpful for drug design, virtual screening and visualizing the chemical space of large and diverse datasets.[18]

In order to obtain the best subset of molecular descriptors in this work, genetic algorithm (GA)[19] was used as an optimization technique. In this respect, GA searches for choosing the best subset of class-specific descriptors, which led to CPANN models with high classification rates. The classification of the molecules in this work was performed according to two different strategies: (1) Classification of the collected molecules based on their therapeutic target, and (2) Classification of compounds according to their activities.

Generally, the present work paves the way for finding specific patterns of molecular descriptors, which are unique for a particular class of anticancer compounds. The general SAR patterns obtained in this work are useful for designing new anticancer molecules and assist the medicinal chemists to screen large databases of compounds, more precisely.

2 Material and Methods

2.1 Dataset

The data in this work consisted of 2904 EGFR-INs, 1072 HD-INs, 1239 MMP-INs and 1074 AR-INs. The molecules of each class of compounds were downloaded as sdf format from the Binding database website. Four sd files were downloaded, each of which was specific for a particular class of anticancer compounds. The sd files have been converted to hin files, using OpenBabel software.[20] This software is an open source toolbox designed for converting different molecular file formats. The obtained hin files contained the structure information of all molecules with a particular therapeutic target. The information of the molecules was provided in these files all together, however for structure optimization, it was necessary to save the structure information of each single molecule, separately. A C^{++} program was written for detaching the information of the molecules and saving them into separate hin files, each contains the structure information of a single molecule. A script was written in Hyperchem software[21] for opening hin files and optimizing the three dimensional (3D) structure of the molecules. The 3D structures of all molecules were optimized by using Austin model 1 Hamiltonian implemented in Hyperchem software. The optimized molecules were used for the calculation of molecular descriptors by using DRAGON software.[22] A total of 1497 0-, 1-, 2-, and 3D descriptors were generated. Descriptors with zero or constant values were eliminated. A total of 416 meaningful molecular descriptors were chosen and considered as independent variables for the model development and further analysis. These variables consisted of constitutional, topological, electronic, geometric and empirical descriptors and encoded different physical and chemical properties of the molecules.

Each sd file contains different activity information for a specific molecule. Such information describes the affinity of a single molecule for different therapeutic targets. A C^{++} program was written in this work for finding particular activity information in an sd file. The activity information for each class of anticancer molecules was extracted from the downloaded sd files and saved in a separate vector.

The structural information of the molecules and calculated molecular descriptors together with their activities are given in Supporting Information section. The detail enumeration of the collected data is given in Table 1. The molecules with IC_{50} and K_i values less than 200 nM were considered as active anticancer agents, while those with IC_{50} and K_i values greater than 2000 nM were considered as inactive ones. The molecules with the activity values within the range of 200 nM–2000 nM were considered as intermediates. The active, inactive and intermediate molecules were used for the development and evaluation of the active-in-

Table 1. Detailed enumeration of the complied anticancer molecules.

	Active			Intermediate		Inactive			Total number of molecules
	Range of activity		Number of compounds	Range of activity	Number of compounds	Range of activity		Number of compounds	
HD-INs	IC50	0–200 nM	560	200 nM < IC50 < 2000 nM	248	IC50	> 2000 nM	255	
	Ki	0–200 nM	6	200 nM < Ki < 2000 nM	1	Ki	> 2000 nM	2	
			566		**249**			**257**	**1072**
AR-INs	IC50	0–200 nM	234	200 nM < IC50 < 2000 nM	223	IC50	> 2000 nM	401	
	Ki	0–200 nM	143	200 nM < Ki < 2000 nM	33	Ki	> 2000 nM	40	
			377		**256**			**441**	**1074**
EGFR-INs	IC50	0–200 nM	1073	200 nM < IC50 < 2000 nM	504	IC50	> 2000 nM	1270	
	Ki	0–200 nM	17	200 nM < Ki < 2000 nM	7	Ki	> 2000 nM	33	
			1090		**511**			**1303**	**2904**
MMP-INs	IC50	0–200 nM	825	200 nM < IC50 < 2000 nM	99	IC50	> 2000 nM	88	
	Ki	0–200 nM	104	200 nM < Ki < 2000 nM	26	Ki	> 2000 nM	97	
			929		**125**			**185**	**1239**
Σ			**2962**		**1141**			**2186**	**6289**

active classifies. These models are useful for screening large databases and finding new potential anticancer molecules.

2.2 Genetic Algorithm-Counterpropagation Artificial Neural Networks

Counterpropagation artificial neural network (CPANN) is a classification and regression technique which combines features from both supervised and unsupervised learnings. This technique is based on the search for object similarities and allows projecting the samples into a topological space, where similar samples are collected in neighbor neurons while dissimilar ones are far apart. The clustered neurons in the map represent combination of different features in the dataset. This property makes CPANN a powerful tool for tackling the classification problems with nonlinear characteristics. Transparent modeling is another advantage of this technique, which makes it less "black box" and the user partially understand what is happening inside the predictive model.

The CPANN architecture consisted of two layers of neurons: a *Kohonen layer* and an *output layer*. The Kohonen layer consists of a grid of Z^2 neurons, where Z is the number of neurons for each side of the toroidal space. Each neuron consists of as many weights as the number of input variables. The weights of neurons are corrected during the training iterations according to the *comparative learning* procedure. At the end of the training procedure, similar objects are collected in neurons with similar weights. Therefore, the weight profile of a specific class of neurons is unique and remarkably different from those of other clusters of neurons. This property is helpful for interpretation of the classification models and is useful for understanding the influences of variables for discriminating different classes of samples.

The performance of the CPANN classifier models depends on different parameters, such as initial and final learning rate, number of training iterations and input subset of variables.[23] In this work, GA has been used for optimizing these parameters and obtaining CPANN models with high classification rates. The proposed GA-CPANN technique is able to uncover the similarities between the molecules and choose the best subset of variables for discriminating different classes of compounds.

GA is an implementation of various search paradigms inspired by natural evolution. This heuristic is routinely used to generate useful solutions for optimization and search problems.[24] At a very general level, GA tries to find the proper values of the variables of a system. The detailed theory of CPANN and GA can be found in literature[16,24a] and their discussion is restrained for the sake of brevity. In this study, GA was used for choosing the appropriate molecular descriptors, proper values of initial and final learning rates and optimum number of iterations for training the network. In this case, GA would search to find the best collection of molecular descriptors and network parameters by optimizing the objective function of the system. Definition of an appropriate objective function as a correct response of the system is an important task for proper determination of the variables. The following objective function has been used in this work for the optimization procedure.

$$\textit{Objective function} = 1-(P + Se + Sp + Sel)/4 \qquad (1)$$

where *P*, *Se*, *Sp* and *Sel* are, respectively correct percent of classification, sensitivity, specificity and selectivity of CPANN models for Venetian blind cross validation ($n=20$). In the present work, the objective function is minimized by using the GA formalism, until the best parameters and descriptors are selected. The GA searches for finding a compromise between the number of interred molecular descriptors and the value of the objective function. After a reasonable promotion of the GA, the subset of molecular descriptors with

Table 2. The specification of the parameters of CPANN and GA used for the optimization procedure.

GA parameters		CPANN parameters	
Number of generation	400	Max. number of training iterations	50
Population size	80	Max. number of nodes in a row	12
Mutation rate	0.2	Neuron bounding	Toroidal
Crossover function	Scattered	Initial weights	Random
Migration condition	Forward	Min. learning rate	0.1
Migration Interval	20	Max. learning rate	0.05

the best discriminative information would be selected for performing the classification. The specification of the GA and CPANN parameters used for optimization in this work are given in Table 2.

2.3 Principal Component Analysis

The PCA algorithm is a popular projection method used to transform an N-dimensional space into an M-dimensional one ($M<N$) created by M uncorrelated vectors called principal components (PCs). This transformation is defined in such a way that the first PC has as high a variance as possible and each succeeding component in turn has the highest variance under the constraint that it be orthogonal to the preceding components. The use of the PCA algorithm is greatly accepted for visualization of large databases. This method is very popular according to its simplicity and underlying methodology for visualization and graphical characterization of high dimensional chemical spaces.[10d, 18]

In order to visualize and interpret the resulting descriptor spaces in this work, the PCA algorithm has been used. This method is applied after the development of the CPANN classifiers and selection of the molecular descriptors. The PCA algorithm was used to visualize the correlation between the target type and representation of the molecules in low dimensional PCA space. Accordingly, the molecules were plotted in the space of the first two or three PCs of the selected molecular descriptors. In this regard we showed that the molecules are clustered in the reduced space of the selected molecular descriptors according to their therapeutic targets and activities.

The MATLAB package (version 7.0.4)[25] was used for running the CPANN (Kohonen toolbox) and genetic algorithm (genetic algorithm toolbox) programs. The calculations in this work were performed using a Pentium 4 personal computer (Intel, Core (TM), Quad, Q9600, and LGA775) with 8 MB of random-access memory (RAM).

3 Results and Discussion

The collected data in this work consists of the main four categories of anticancer molecules (AR-INs, HD-INs, MMP-INs, and EGFR-INs). In order to generate the classification models the data was divided into three groups of active, intermediate and inactive molecules. The criteria for this division and the distribution of compounds into these groups are summarized in Table 1.

As mentioned in previous sections, the classification was performed according to two different strategies: (1) Classification of compounds according to their therapeutic target. In this respect, the active anticancer molecules were classified into four groups of AR-INs, HD-INs, MMP-INs, and EGFR-INs and the best class-specific descriptors were identified for the each class of compounds. Classification of small molecules according to their target type is one of the main subjects in chemoinformatics.[18b] (2) Classification based on activity of compounds. In this case, the binary classifiers were developed for discriminating between the active and inactive molecules in each class of the anticancer compounds. The active-inactive classifier models contain descriptors which code information about the activity of a specific class of compounds. The values of these molecular descriptors show different patterns for active and inactive molecules.

Only active molecules (see Table 1) were used for the classification of compounds according to their therapeutic target. The reason is that only active molecules contain information about the ligand-target interactions and inactive ones are very poor from this point of view. The flowchart of classification procedure is presented in Figure 1.

As shown in Figure 1, two different approaches were implemented for deriving the active-inactive classifiers: (1) Development of active-inactive binary classifiers by using the active and inactive molecules (see Table 1). In this case an activity gap exists between the two groups of molecules. (2) Development of the binary active-inactive classifiers by using the active, inactive and intermediate molecules. Here, the intermediate molecules were also considered as "active" objects for training the models. The accuracy of the predictions of the developed models for both approaches was compared. The followings describe the results obtained for the classification of compounds based on their therapeutic target and activities.

3.1 Classification of Compound Based on Their Therapeutic Targets

The main aim of this section was to develop the classification models, which are able to discriminate between compounds according to their therapeutic target. These models should be able to map the input space of molecular de-

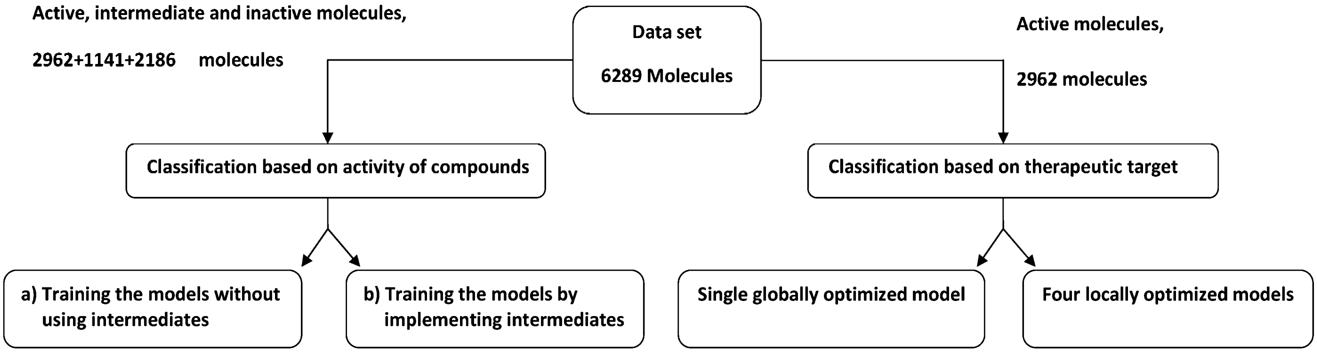
Active, intermediate and inactive molecules,
2962+1141+2186 molecules
Data set
6289 Molecules
Active molecules,
2962 molecules
Classification based on activity of compounds
Classification based on therapeutic target
a) Training the models without using intermediates
b) Training the models by implementing intermediates
Single globally optimized model
Four locally optimized models

Figure 1. Flowchart of the classification procedure.

scriptors to an output space of more than two pattern classes. Four groups of anticancer molecules with different targets are considered in this work, therefore their classification is a problem of multi-group classification. There are two basic approaches to the multi-group classification problem.[26] The first is using a single globally optimized model to discriminate among all classes. The training process of a single multi-class classifier is easy to handle, however in this process all pattern classes should use the same feature space. The second approach to multi-group classification is development of multiple locally optimized models to perform different pieces of the discrimination task. Since multiple locally optimized classifiers are trained independently, each model has its own feature space and a special feature extraction can be designed to fit each classifier model. A system of multiple binary classifier models provides a lot of flexibility compared to those of a single globally optimized model.[27] In this work, both approaches of multigroup classification are implemented for the classification of active anticancer molecules.

3.1.1 Multiclass Classification Using a Single Globally Optimized Model

A collection of 2962 active anticancer molecules consisted of 929 MMP-INs, 377 AR-INs, 1090 EGF-INs and 566 HD-INs was selected from the data. 15 % of the active molecules in each class were selected randomly, as a test set. The remaining molecules were used for the training procedure. All of the modeling steps (selection of the molecular descriptors, determination of the size of the networks and the weights of the neurons) were performed by using the molecules of the training set. In this case the objective function (Equation 1) was minimized by GA, until the best subset of molecular descriptors, network size and learning rate were selected. The molecules of the test set did not influence the model selection procedure and also had no role for the determination of the parameters of the CPANN models.

The initial descriptor pool consisted of 416 molecular descriptors. GA-CPANN algorithm was used for the selection of the best molecular descriptors and optimization of the network parameters, simultaneously. The selected molecular descriptors and statistical parameters of the developed CPANN model are presented in Table 3. The confusion matrix of the CPANN model is also provided in Table 4. As can be seen in these tables the developed CPANN model can correctly classify the molecules according to their inhibition behavior. This model contains six molecular descriptors of mean electrotopological state (MES), molecular local dipole index (LDI), molecular refractivity index (MRI), mean square distance index (MSD) and the number of X—CH—X and X—CH..X fragments. The parameter of X—CH—X is defined as a central carbon atom on an aromatic ring that has two hetroatoms as neighbor on the same aromatic ring and the third neighbor outside this ring is a hydrogen. The descriptor of X—CH..X has the same meaning however, '..' stands for an aromatic single bond. By considering the selected molecular properties, one may understand the inhibition behavior of the molecules. The trained Kohonen network together with the label of each molecule in each neuron is demonstrated in Figs. S1a–S1d in Supporting Information section. Figs. S1a–S1d are colored according to the class membership values of neurons for AR-INs, EGFR-INs, HD-INs and MMP-INs, respectively. The Kohonen map of anticancer molecules is a visual representation of the selected chemical space and compounds are distributed in this map according to their activity toward a particular therapeutic target. Using Kohonen networks and SOM maps for visualizing the chemical space is a common procedure in virtual screening and drug discovery projects.

As mentioned in the previous section, the weight profiles of the *Kohonen layer* are useful for understanding the classification model. The average values of the *Kohonen weights* of the developed model, for each class and each selected molecular descriptor, are shown in Figure 2. The discriminative power of the selected molecular descriptors and their interactions can be deduced from this figure. As can be seen, the main effects of MSD, MRI, X—CH—X and X—CH..X parameters are discriminative and represent different patterns for different classes of molecules. The bar plots of the values of the parameters of MSD, MRI, X—CH—X and X—CH..X for four classes of active anticancer molecules are shown in Figures S2a–S2d, respectively.

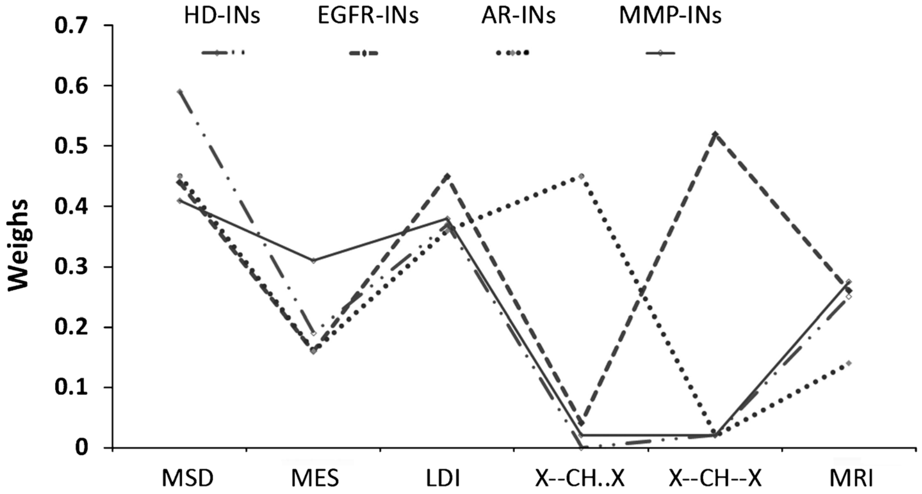

Figure 2. Average values of the weights of the CPANN model, for each class and each selected molecular descriptor.

Inspection of Figure S2a reveals that the values of MSD variable for HD-INs are very different from those of the other three classes of compounds. The MSD values of HD-INs are significantly higher than those of other classes of anticancer molecules. MSD is the mean square distance index and is calculated from the second order distance distribution moments of the *H-depleted* molecular graph. This parameter decreases with increasing the molecular branching.[28]

As can be seen in Figure S2b, the values of MRI for AR-INs are meaningfully lower than those of other classes of active anticancer molecules. MRI contains information about the volume and polarizability of the molecules. This descriptor is related to the London dispersive forces acting in the drug-receptor interactions.[29]

Figures S2c and S2d show that the X–CH–X and X–CH..X fragments are specific indicators for EGFR-INs and AR-INs, respectively. Both X–CH–X and X–CH..X are among atom centered fragment (ACF) descriptors. ACF descriptors are very helpful references for screening large databases of compounds and it is shown that they are useful for similarity analysis and virtual screening studies.[30]

The discriminative power of the interaction of the LDI and MES for classifying the active anticancer molecules is shown in Figure 3. As can be seen, by considering the values of these molecular descriptors one may be able to distinguish the MMP-INs from the HD-INs, EGFR-INs and AR-INs. Inspection of this figure reveals that for a given value of LDI, MMP-INs show higher MES values and *vice versa*. It can be concluded that the ratio of the MES to LDI, (*MES/LDI*), is higher for the MMP-INs compared to those of the HD-INs, EGFR-INs and AR-INs. The MES is the mean value of the electrotopological state (E-state) indices of all non-hydrogen atoms in a molecule. The E-state index of an atom gives the information related to the ratio of π and lone pair electrons to the number of the σ bonds in the molecular graph for the considered atom.[31] The large and positive value of the MES shows that the molecule is composed of the electronegative atoms or atoms with high electronic accessibility. These compounds consisted of atoms with high probability of interaction with the other molecules. Low value of MES represents that the compound mainly consists of atoms possessing only σ electrons.[28] LDI is the average of the charge differences over all *i–j* bonded atom pairs in a molecule.[32] This parameter measures the amount of local dipole moments in a molecule. Considering the definition of the MES and LDI, the MMP-INs show higher electronic accessibility and lower local dipole moments compared to the HD-INs, EGFR-INs and AR-INs.

The selected six molecular descriptors in this section define discrete areas in the existing chemical space for particular classes of anticancer molecules. The projections of molecules into the first three principal components of the selected molecular descriptors are illustrated in Figure 4. By considering this figure one can characterize discrete areas for different classes of anticancer molecules in the space of the principal components.

3.1.2 Multiclass Classification Using Multiple Locally Optimized Models

In order to classify each group of anticancer compounds independently, four one-class classifiers were developed. Each of these models should be able to distinguish a specific class of anticancer compounds from the other three classes. The GA-CPANN algorithm was used for developing four different CPANN models, each one just discerns a specific group of molecules from the remaining ones. The GA-CPANN algorithm was run four times, in each case with a specific output vector. The elements of output vector for the molecules of the class of interest were assigned to be

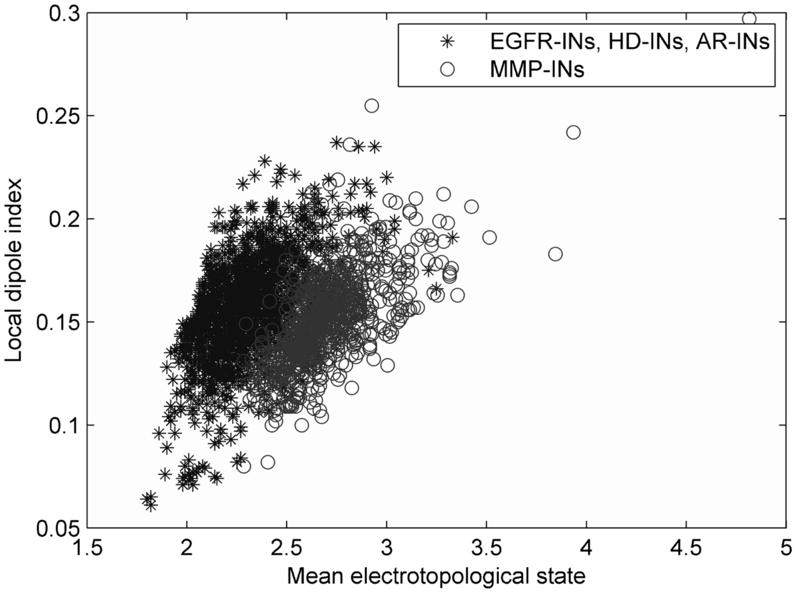

Figure 3. Local dipole index (*LDI*) versus mean electrotopological state (*MES*) for different classes of active anticancer compounds.

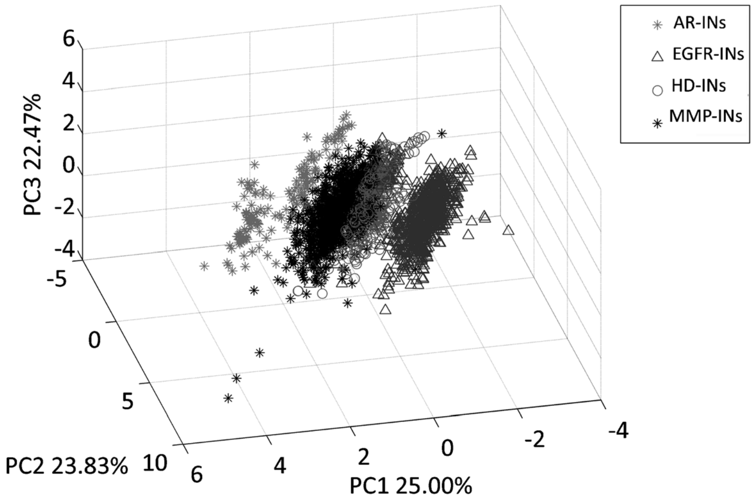

Figure 4. Projections of active anticancer molecules into the first three principal components of the selected molecular descriptors.

one, while for the remaining ones these elements were set to be zero. The statistical results for the training and test sets together with the selected variables for the four developed CPANN models are given in Table 3. The detailed description of the selected molecular descriptors and their definition can be found in literature.[28] These models are accurate and their molecular descriptors contain information, which is solely specific for determining the inhibition behavior of a specific class of molecules. These binary classifiers can subsequently be used to predict the probability of new compounds to be active against a given biological target.

Both single and locally optimized classifiers in this work are useful for discriminating between different active anticancer molecules and determine the molecular properties, which are specific indicators for a particular class of anticancer molecules.

3.2 Classification Based on Activity of Compounds

3.2.1 Development of the CPANN Models

As previously mentioned, two different approaches were implemented for the classification of molecules according to their activities. We were interested to see which strategy is better in predicting active, inactive and intermediate molecules. At first, the binary active-inactive classifiers were built by using the active and inactive molecules. In this case, an activity gap exists between the active and inactive groups (see Table 1). 15% of active and inactive molecules

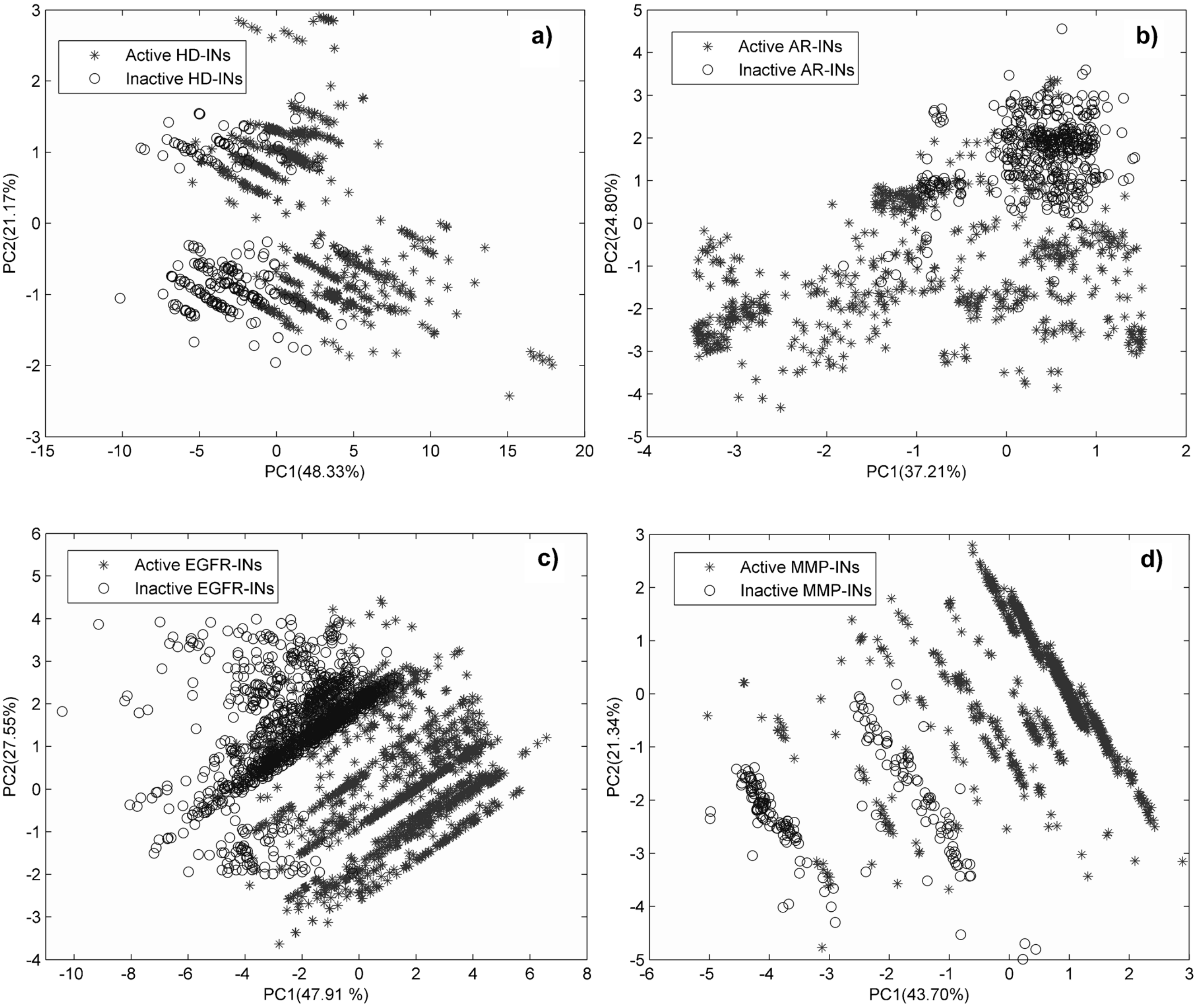

Figure 5. Projections of active and inactive molecules to the first two principal components of the selected molecular descriptors. a) HD-INs, b) AR-INs, c) EGFR-INs, d) MMP-INs. The active and intermediate molecules were considered in the same group, ("active") in this figure.

Table 3. The variables and statistical parameters of the CPANN models for classification of the active anticancer compounds.

	Selected molecular descriptors by GA-CPANN	Specifity/Sensitivity/ Precision	Correct percent of classification	
			Training set	Test set
Global optimized model	Electrotopological state index, Local dipole index, Molecular refractivity index, mean square distance index, number of X–CH–X and X–CH..X fragments	0.942/0.953/0.947	93.36 %	91.16 %
Local model for HD-INs	Maximum positive charge, Number of halogen atoms, Electrotopological variation, Number of-CONHRPh groups	0.946/0.987/0.988	91.18 %	90.12 %
Local model for AR-INs	Number of X–CH..X fragments, Number of acceptor atoms for H-bonding, Number of CR4 fragments	0.978/0.955/0.985	96.85 %	92.35 %
Local model for EGFR-INs	Number of X–CH–X fragments, number of –NHRPh groups, Bond-order ID number	0. 989/0.989/0.987	98.09 %	94.22 %
Local model for MMP-INs	Number of R-SO2-R groups, Maximum positive charge, Sum of topological distance between N and P (T(N..P))	0.943/0.944/0.948	94.11 %	92.73 %

Table 4. The confusion matrix of the CPANN model used for classification of the molecules of the training and prediction sets. HD-INs: Class 1, AR-INs: Class2, EGFR-INs: Class 3, MMP-INs: Class 4.

Predicted/Real	Training set					Test set			
	Class 1	Class 2	Class 3	Class 4		Class 1	Class2	Class 3	Class4
Class 1	**467**	7	2	7	Class 1	**78**	2	5	1
Class 2	5	**264**	8	12	Class 2	2	**48**	0	6
Class 3	4	14	**905**	37	Class 3	1	3	**157**	1
Class4	5	35	11	**733**	Class4	4	4	2	**132**

in each class were randomly selected as test set. The remaining molecules were used for the training procedure. The test set molecules had no effect on the model selection procedure and had no role in determining the parameters of the classifiers. Four CPANN models were constructed by using the training molecules. The GA-CPANN algorithm was run four times in each case with active and inactive molecules of a particular class of compounds as training set. Each CPANN model was designed to discriminate active molecules of a particular class of compounds from the inactive ones. After the development of the models, the test set was used for the evaluation of the performances of the classifiers. In addition, 15% of the intermediate molecules were randomly selected and predicted by the developed models. These compounds were supposed to be predicted as "inactive", because they are above the active threshold.

As another approach, the intermediate molecules were also used for training the models. The splitting of the molecules was the same as those of the previous approach, except that the 85% of the intermediate molecules in each class were also used for training the models. The elements of the output vector for the active and intermediate molecules were set to be "1" and for the inactive molecules were set to be zero. The intermediate molecules were considered as "active" objects for training the models. The classification performances of the developed models by using the two described approaches are given in Table 5. As can be seen in this table, the performances of the models in both approaches are comparable for predicting the active and inactive molecules. However the main difference is the prediction of the intermediates. The developed models by the second approach show better performances for predicting the test set molecules of the intermediates. The results in this section reveal the influences of the intermediate molecules on the performances of the binary active-inactive classifiers. As can be seen in Table 5, implementation of the intermediate molecules for model training enhanced the performances of the classifiers for predicting the activity level of the molecules. It can be concluded that the intermediate molecules in this work contain important hits and should be considered in "active" group for training the binary active-inactive classifiers.

The selected molecular descriptors of the binary classifiers, developed by implementing the intermediate molecules in "active" group, are given in Table 6. The majority of the selected molecular descriptors are among atom centered fragments and molecular indicators and measure the number of specific group of atoms in a molecule (see Table 6). Such indicators are very informative indices for describing the interaction of the molecules with their biological targets and can be used as virtual filters for the screening of large databases of compounds, such as ZINC[33] collection. More description about the selected molecular descriptors can be found in literature.[28] The mean values of the selected descriptors are expected to significantly vary from active to inactive molecules. In order to examine this hypothesis, student's t-test[34] was used. The results are summarized in Table 6. According to the calculated *p* values, the mean values of the selected molecular descriptors were found to be significantly different for the active and inactive molecules. Therefore, these variables can be considered as molecular filters for the screening of compound databases and help the medicinal chemists to speed up the screening stage of the drug discovery projects.

3.2.2 Estimating the Prediction Reliability of the Models

The results of the previous section showed the superiority of the models trained by implementing the intermediate molecules. This section focuses on the estimation of the

Table 5. The correct percent of classifications for the developed binary active-inactive classifiers.

	Model trained with active and inactive molecules						Model trained with active, inactive and intermediates					
	Active		Intermediate		Inactive		Active		Intermediate		Inactive	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Model 1[a]	87.52%	87.05%	–	56.75%	83.03%	79.48%	86.48%	78.82%	87.73%	81.08%	81.65%	82.05%
Model 2	85.62%	84.21%	–	47.36%	91.73%	81.81%	83.75%	78.94%	88.07%	86.84%	85.33%	86.36%
Model 3	91.68%	84.14%	–	66.14%	87.00%	78.06%	88.55%	81.70%	85.25%	87.01%	87.62%	85.71%
Model 4	82.40%	84.90%	–	42.10%	93.00%	89.28%	83.54%	82.01%	83.01%	89.47%	92.35%	96.42%

[a] Model 1: HD-INs; Model 2: AR-INs; Model 3: EGFR-INs; Model 4: MMP-INs.

Table 6. The selected molecular descriptors for classifying anticancer compounds into two groups of active and inactive.

Active-Inactive classifier	Selected molecular descriptors	Mean value for active [a] molecules	Mean value for inactive molecules	Calculated *p* value
HD-INs	Relative positive charge (RPCG)	0.113	0.145	0.000
	Number of fluorine atoms attached to hetro atom	0.045	0.513	0.000
	Number of R=S fragments	0.019	0.170	0.000
	Topological charge index of order 1	1.113	0.903	0.000
	Number of –CONHRPh groups	0.555	0.134	0.000
	Narumi simple topological index	22.82	18.16	0.000
	Total path count	2128	1188	0.000
AR-INs	Number of cyanide groups	0.386	0.013	0.000
	Number of X–CH..X fragments	0.858	0.165	0.000
	Number of aromatic bonds	13.05	9.820	0.000
	Number of sulfur atoms	0.232	0.095	0.000
	Number of =CR2 fragments	0.073	0.268	0.000
	Quadratic index	19.24	16.37	0.000
	Spanning tree number	5.915	5.080	0.000
EGFR-INs	Eccentric index (DECC)	1.972	1.571	0.000
	Number of non-terminal carbon (SP)	0.284	0.018	0.000
	Number of phenyl groups	0.043	0.337	0.000
	Number of X–CX–X fragments	0.066	0.357	0.000
	Graph vertex complexity index	3.705	3.174	0.000
	Sum of topological distance between N and Cl	13.11	5.657	0.000
	Average geometry distance degree	338.0	266.7	0.000
MMP-INs	Number of ten membered rings	0.076	0.550	0.000
	Naurimi harmonic topological index	1.631	1.873	0.000
	Number of R–CR–X fragments	1.074	0.179	0.000
	Ratio of multiple path count to path count (PCR)	15.41	25.92	0.000
	Number of sulfonamides	0.717	0.163	0.000
	Asphericity	0.332	0.547	0.000

[a] The active and intermediate molecules were considered in the same group ("active").

prediction reliability of these models and the projection of the active and inactive molecules on the first two PCs of the selected molecular descriptors.

In order to have a general insight into the reliability of the predictions, the applicability domains (*AD*) of the models were calculated. There are several methods for estimating the applicability domain of the in-silico models.[35] In the present work the k-nearest neighbor (*k*NN) approach, based on Elucidation distance in descriptor space, was utilized for determination of the *AD* of the binary active-inactive classifiers. The *AD* was calculated from the distribution of the similarities between each compound and its *k*-nearest neighbor in the training set. The similarities are calculated using average Elucidation distance of *k*-nearest neighbor of a molecule in the training set using only a subset of molecular descriptors identified by the modeling procedure as optimal. The distribution of distances between each molecule and its nearest neighbor in the training set is computed to produce an applicability domain threshold (AD_T), as follows:

$$AD_T = \bar{D} + Z\sigma$$

where $\bar{D}$ is the average Elucidation distance of the *k*-nearest neighbors of compounds within the training set, σ is the standard deviation of Elucidation distances, and Z is an arbitrary parameter to control the significant level. In this work Z is equal to 0.5. If the distance of a test compound from any of its k-nearest neighbors in the training set exceeds the threshold, the prediction is considered as unreliable. Calculations within descriptor space in this work were initially assessed using the Elucidation distance with selected descriptors being mean centered and autoscaled. Table 7 summarizes the calculated AD_T of the developed binary active-inactive classifiers in this work, together with the number of "in domain" test set molecules. As can be seen in this table, the most of the test set molecules fall within the *AD* of the models and therefore the majority of the predictions in this work are reliable.

The projections of active and inactive molecules to the first two principal components of the selected molecular descriptors, for the four classes of anticancer molecules are shown in Figures 5a–d. As can be seen in these figures, active and inactive molecules are reasonably distinguishable by considering the values of first and second principal components. It can be concluded that the selected molecu-

Table 7. The calculated applicability domain (*AD*) for the developed binary active-inactive classifiers.

	AD_T (k [a] = 5)	Number of "in domain" test set molecules	AD_T (k [a] = 10)	Number of "in domain" test set molecules
CPANN model for HD-INs	0.58	147/161 (91.30%)	0.69	151/161 (93.78%)
CPANN model for AR-INs	0.62	150/161 (93.16%)	0.74	154/161 (95.62%)
CPANN model for EGFR-INs	0.54	412/436 (94.50%)	0.58	413/436 (94.72%)
CPANN model for MMP-INs	0.72	172/186 (92.47%)	0.82	177/186 (95.16%)

[a] Number of nearest neighbors

lar descriptors can describe discrete areas in chemical space for the active and inactive anticancer molecules.

The developed active-inactive classifiers in this work are trained with structurally diverse sets of molecules and significant number of anticancer compounds. Therefore, these models are general and robust and can actually speed up the projects dealing with the discovery of new anticancer molecules.

4 Conclusions

The main aim of the present contribution was to develop general SAR rules and classification patterns for anticancer molecules. Such rules and patterns can be considered as virtual filters for mining of the large databases of compounds and finding new anticancer candidates. A collection of 6289 anticancer molecules with known experimental activities was collected from Binding database and was used for the development of the models. The method of counterpropagation artificial neural network was used for the classification of the data and the results revealed that the compounds with similar bioactivities tend to cluster in the space of the selected molecular descriptors. The classification was performed according to two different routes: (1) Classification of compounds based on their therapeutic target. (2) Classification based on activities. Both strategies led to some simple and effective SAR patterns, which can be summarized as follows: (1) The values of the refractive index for the AR-INs are meaningfully less than those of the HD-INs, MMP-INs and EGFR-INs. (2) The ratio of the MES to LDI, (*MES/LDI*), is higher for the MMP-INs compared to those of HD-INs, EGFR-INs and AR-INs (3) X–CH..X fragments is a useful indicator for the AR-INs. (4) Number of aromatic bonds for the active AR-INs is meaningfully greater than those of the inactive AR-INs. (5) The MSD index for HD-INs is greater than those of the EGFR-INs, MMP-INs and AR-INs. (6) The values of RPCG and number of –R=S fragments for the active HD-INs are meaningfully less than those of the inactive HD-INs.

Generally, the collected data and the developed classifier models in this work could represent a powerful tool for designing novel pharmacological agents and allow the analysis of large databases with the aim of discovering new compounds with potent anticancer activity.

Supporting Information Available

The calculated molecular descriptors and SD file format of the molecules, together with Figures S1a–S1d and Figures S2a–S2d, are given in Supporting Information.

References

[1] American Cancer Society, www.cencer.org (accessed **2011**).

[2] A. Jemal, A. Seigel, J. Xu, E. Ward, *CA Cancer J. Clin.* **2010**, *60*, 277–300.

[3] S. S. Cross, Curr. Diagn. *Pathology* **2005**, *11*, 329–339.

[4] R. D. Rubens, R. E. Coleman, *Cancer Treat. Rev.* **1999**, *25*, 1–2.

[5] H. Ovaa, C. Kuijl, J. Neefjes, *Drug Discov. Today* **2009**, *6*, 1–4.

[6] a) http://www.ariad.com (accessed **2011**); b) http://www.adventrx.com (accessed 2011); c) http://www.telik.com (accessed **2011**).

[7] H. Kubinyi, *Perspect. Drug Discov. Today* **1998**, *9*, 225–252.

[8] a) C. Lipinski, A. Hopkins, *Nature* **2004**, *432*, 855–861; b) T. I. Opera, J. Gottfries, *J. Comb. Chem.* **2001**, *3*, 157–166; c) D. J. Triggle, *Biochem. Pharm.* **2009**, *78*, 217–223; d) R. Bade, H. Chan, J. Reynisson, *Eur. J. Med. Chem.* **2010**, *45*, 5646–5652.

[9] J. Bajorath, *Nat. Rev. Drug Discov.* **2002**, *1*, 882–894.

[10] a) P. Willett, V. Winterman, D. Bawden, *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109–118; b) G. M. Downs, M. J. Barnard, *Rev. Comput. Chem.* **2002**, *18*, 1–40; c) H. Matter, *J. Med. Chem.* **1997**, *40*, 1219–1229; d) R. V. Deursen, L. C. Blum, J. Reymond, *J. Chem. Inf. Model.* **2010**, *50*, 1924–1934.

[11] T. Liu, Y. Lin, X. Wen, R. N. Jorrisen, M. K. Gilson, *Nucleic Acids Res.* **2007**, *35*, 190–201; b) X. Chen, Y. Lin, M. K. Gilson, *Nucleic Acid Sci.* **2002**, *61*, 127–141; c) X. Chen, Y. Lin, M. Liu, M. K. Gilson, *Bioinformatics* **2002**, *18*, 130–139; d) X. Chen, M. Liu, M. K. Gilson, *J. Comb. Chem.* **2001**, *4*, 719–725; e) http://www.bindingdb.org (accessed **2011**).

[12] R. Somech, S. Izraeli, A. J. Simon, *Cancer Treat. Rev.* **2004**, *30*, 461–472.

[13] A. Brodie, *Trends Endocrin. Met.* **2002**, *13*, 61–65.

[14] A. Wakeling, *Curr. Opin. Pharmacol.* **2002**, *2*, 382–387.

[15] M. Seiki, *Cancer Lett.* **2003**, *194*, 1–11.

[16] D. Ballabio, V. Consonni, R. Todeschini, *Chemom. Intel. Lab. Syst.* **2009**, *98*, 115–122.

[17] T. Kohonen, *Self-organizing Maps*, 3rd ed, Springer, Berlin, **2001**.

[18] a) D. Hiristozov, T. I. Oprea, J. Gasteiger, *J. Chem. Inf. Model.* **2007**, *47*, 2044–2062; b) M. V. Korrf, K. Hilpert, *J. Chem. Inf. Model.* **2006**, *46*, 1580–1587.

[19] L. Terfloth, J. Gasteiger, *Drug Discov. Today* **2001**, *6*, 102–108.

[20] *OpenBabel: The open source chemistry toolbox*, http://openbabel.org/wiki/main_page (accessed **2011**).

[21] *HyperChem*, Hypercube, Inc. http://www.hyper.com (accessed **2011**).

[22] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, *Dragon Software: Calculation of Molecular Descriptors,* Department of Environmental Sciences, University of Milano- Bicocca, and Talete, srl. http://disat.unimib.it/chm/Dragon.htm, Milan, Italy, **2003**.
[23] W. Melssen, R. Wehrens, L. Buydens, *Chemom. Intel. Lab. Syst.* **2006**, *83*, 99–113.
[24] a) R. Leardi, *J. Chromatogr. A* **2007**, *1185*, 226–233; b) V. Venkatasubramanian, K. Chan, J. M. Caruthers, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 188–195.
[25] MATLAB 7.4, http://www.mathworks.com/products/matlab (Accessed **2011**).
[26] R. Brereton, *Chemometrics for Pattern Recognition*, Wiley, Chichester, West Sussex, UK, **2009**.
[27] G. Ou, Y. Murphey, *Pattern Recogn.* **2007**, *40*, 4–18.
[28] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors, Methods and Principles in Medicinal Chemistry*, Wiley-VCH, Weinheim, **2000**.
[29] A. Padrón, R. Carrascol, R. F. Pellón, *J. Pharm. Pharmaceut. Sci.* **2002**, *5*, 258–266.
[30] a) R. Kuhne, R. Ebert, G. Schuurmann, *J. Chem. Inf. Model.* **2009**, *49*, 2660–2669; b) J. Batista, L. Tan, J. Bajorath, *J. Chem. Inf. Model.* **2010**, *50*, 79–86.
[31] L. H. Hall, B. Mohney, L. B. Kier, *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76–82.
[32] M. Karelson, V. S. Lobanov, *Chem. Rev.* **1996**, *96*, 1027–1043.
[33] *ZINC database, a free database of commercially-available compounds for virtual screening*, http://zinc.docking.org/ (Accessed **2011**).
[34] S. N. Deming, S. L. Morgan, *Experimental Design: A Chemometric Approach*, Elsevier Science Publisher, Amsterdam, **1994**.
[35] S. Weaver, M. P. Gleeson, *J. Mol. Graph. Model.* **2008**, *26*, 1315–1326.

Received: June 16, 2011
Accepted: November 30, 2011
Published online: January 13, 2012