The Importance of Descriptor-Based Clusterization in QSAR Models Development: Tyrosine Kinases Inhibitors as a Key Study

ARTICLE *in* MOLECULAR INFORMATICS · AUGUST 2011

Impact Factor: 1.65 · DOI: 10.1002/minf.201100036

Giovanni Marzaro
University of Padova
30 PUBLICATIONS 259 CITATIONS

Paola Brun
University of Padova
102 PUBLICATIONS 1,494 CITATIONS

Ignazio Castagliuolo
University of Padova
110 PUBLICATIONS 2,458 CITATIONS

Adriana Chilin
University of Padova
89 PUBLICATIONS 764 CITATIONS

DOI: 10.1002/minf.201100036

The Importance of Descriptor-Based Clusterization in QSAR Models Development: Tyrosine Kinases Inhibitors as a Key Study

Giovanni Marzaro,*[a] Francesca Tonus,[a] Paola Brun,[b] Ignazio Castagliuolo,[b] Adriano Guiotto,[a] and Adriana Chilin[a]

Abstract: Quantitative Structure Activity Relationship (QSAR) is a well known cheminformatic tool for the discovery of novel biologically active compounds. However, when large and heterogeneous datasets are mined, it is not possible to derive a QSAR equation able to predict in a satisfactory manner the activity of the compounds. Thus, QSAR models are often inadequate for virtual screening purpose. Herein we present a novel approach to multitarget classification QSAR models, useful to assess the selectivity profile of the tyrosine kinases inhibitors. A descriptor-based clusterization process was employed, that allowed the generation of models with high accuracies and independent from the chemical classification of the compounds (i.e. from the scaffold type). The herein proposed methodology can lead to QSAR models useful for virtual screening processes.

Keywords: QSAR · Clusterization · Tyrosine kinase inhibitors · Quinazolines

1 Introduction

Coined in 1964,[1] the term Quantitative Structure Activity Relationship (QSAR) indicates a well established cheminformatic tool in medicinal chemistry.[2] The QSAR theory is fundamentally based on three assumptions: first, biological properties are related to molecular structures; second, molecular structures can be converted in numerical parameters called descriptors; third, it is possible to find a statistical correlation among biological properties and molecular descriptors. During the years, several strategies to build up QSAR models have been developed, making use of novel molecular descriptors,[3] approaches for descriptors selection (e.g. genetic algorithms[4]) and methods to reduce the dimensionality of the data (e.g. principal component analysis, PCA, or partial least square regression, PLSR[5]). More recently, innovative theoretical approaches (e.g. 3D-QSAR, Hologram-QSAR and machine learning tools[6]) have been introduced. For drug discovery purposes, there are mainly two types of QSAR models: the classification analysis, in which the molecules are considered as active or inactive at a given activity cut-off, and the regression analysis, in which the exact value of the biological activity (generally the pIC_{50}, the association or the dissociation constants) is considered. One of the major challenges performing a QSAR analysis is to derive a model able to predict the activity of (or to classify) structurally heterogeneous molecules. In fact, in order to be useful for virtual screening (VS) processes, QSAR models have to be able to mine data from large and non homogeneous dataset. Unfortunately, this is often an unreachable goal: it is quite obvious that structurally different molecules could show so different interaction geometries with the target that no single one correlation among biological properties and molecular descriptors exists. In other words, when large and heterogeneous datasets are mined, it is not possible to derive a QSAR equation able to predict in a satisfactory manner the activity of the compounds. A useful strategy to enhance QSAR accuracy in nonhomogeneous datasets consists in regrouping the compounds in more homogeneous subsets.[7,8] The clustering algorithms make use of diverse parameters, such as similarity indexes, molecular fingerprints or molecular descriptors. Regrouping the compounds on the basis of structural features can lead to QSAR models with high accuracy. However, these models do not find high relevance in VS, in which large and heterogeneous datasets are mined. In fact, in order to find structurally novel (and thus patentable) drug candidates, a tool for VS has to be as less "structure dependent" as possible. This is of fundamental importance if con-

[a] G. Marzaro, F. Tonus, A. Guiotto, A. Chilin
Department of Pharmaceutical Sciences, University of Padova
Via Marzolo 5, 35131 Padova, Italy
*e-mail: giovanni.marzaro@unipd.it

[b] P. Brun, I. Castagliuolo
Department of Histology, Microbiology and Medical
Biotechnology, University of Padova
Via Gabelli 63, 35121 Padova, Italy

Supporting Information for this article is available on the WWW under http://dx.doi.org/10.1002/minf.201100036.

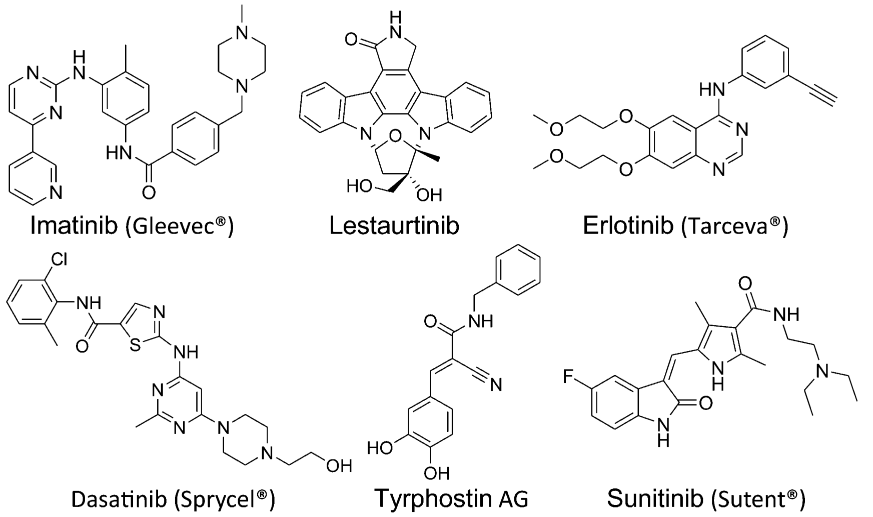

Figure 1. Examples of known TKIs.

sidering that the activity of a target can be modulated by several classes of compounds. In other words, a "*pharmacologically homogeneous*" family of compounds is often a "*chemically non homogeneous*" family. An outstanding example of such a situation is constituted by the tyrosine kinases (TKs) inhibitors (TKIs). TKIs act as ATP-competitive inhibitors and are useful drugs for anticancer therapy.[9] As depicted in Figure 1, the TKIs approved by the FDA for tumor treatment or at least entered in clinical trials show a low degree of structural similarity.[10–15]

Several TKIs show a wide spectrum of activity, as they are able to inhibit a number of TKs.[16] This is of fundamental importance in cancer therapy because it has been recently established that the use of compounds with multi-target activity significantly reduces the drug resistance phenomena onset.[17] A useful strategy to identify novel promising multi-tyrosine kinase inhibitors would be the employment of QSAR models able to predict the activity of novel compounds towards a number of TKs.

Due to our interest in novel lead compounds discovery, we have recently developed a multi-target classification model for TKIs employing a classical QSAR analysis.[18] In this paper we face a different approach, demonstrating the importance of the *descriptor-based* clusterization in a large and heterogeneous dataset, even if it was possible to derive a single QSAR model with good accuracy. In order to reach this goal, we took into consideration a set of previously reported TKIs assayed against at least one of eight TKs involved in cancer disease. Then, several classification multitarget QSAR analyses have been performed. All the models have been evaluated through external validation and through the classification of novel compounds.

2 Results and Discussion

2.1 Datasets

The structural and biological data for the 1359 TKIs employed in the present work were retrieved from literature.[19] The entire dataset have been split into two distinct subsets, called the Main Dataset (MD) and the Virtual Screening Test Set (VSTS). Only the MD has been used to derive the QSAR models. The VSTS, in fact, was constructed so that it contained compounds quite structurally unrelated with those of the MD. The VSTS was used to assess the performances of the QSAR models in predicting the selectivity profile of compounds highly different from those employed to generate the models.

By means of visual structural inspection, the MD and the VSTS resulted altogether constituted by 26 classes of compounds (Figure 2). Only the scaffolds number 2 and 22 were represented in both the MD and the VSTS, while scaffolds 24–26 were present only in the VSTS. Several statistics regarding the cases distribution for each TK have been reported in Table 1.

2.2 Cluster Analysis and QSAR Modeling

In QSAR applications, the cluster analysis is based on the *similar properties principles*, which states that "compounds with similar structure are likely to exhibit similar properties". The purpose of the present work is to demonstrate that "compounds with *some* similar molecular descriptors are likely to obey the same structure-activity relationship". Previously unreported results obtained in our research group revealed that, among the 3224 molecular descriptors computable by Dragon 5.5 software,[20] LP1 and BELp1 descriptors were useful in recognizing not so obvious molecu-

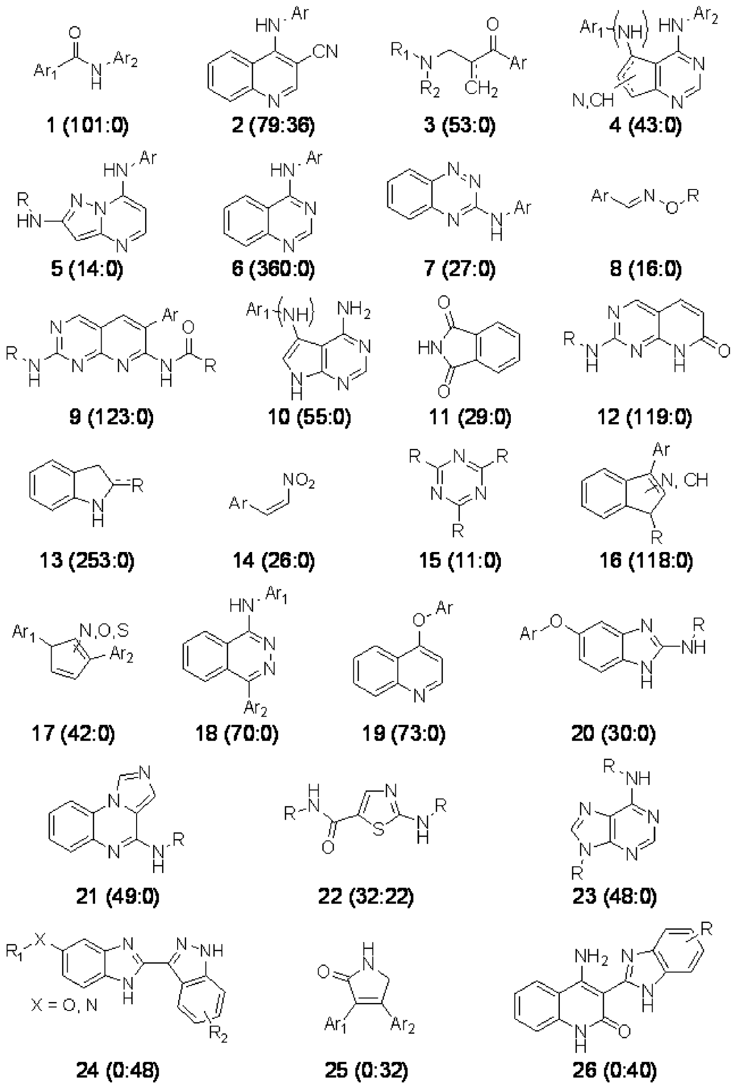

Figure 2. Classes of compounds considered. The MD:VSTS cases ratio are reported between brackets.

Table 1. MD and VSTS statistics.

	Main dataset (MD)			Virtual screening test set (VSTS)		
TK	Number of cases	Active cases at 1.0 μM (%)[a]	Active cases at 0.1 μM (%) [b]	Number of cases	Active cases at 1.0 μM (%) [a]	Active cases at 0.1 μM (%) [b]
abl	102	24.5	18.6	0	–	–
lck	91	91.2	65.9	0	–	–
src	434	65.2	52.1	20	45.0	15.0
EGFR	457	66.1	54.9	17	70.6	35.3
FGFR-1	98	71.4	30.6	22	100.0	81.8
KDR	274	77.4	48.2	60	66.7	35.0
PDGFRβ	240	16.3	14.2	26	69.2	42.3
VEGFR-1	75	37.3	1.3	33	90.9	60.6
TOTAL	**1771**	**58.8**	**42.5**	**178**	**73.6**	**44.4**

[a] Percentage of compounds which IC_{50} was at most 1.0 μM; [b] Percentage of compounds which IC_{50} was at most 0.1 μM.

lar similarities. The LP1 descriptor (Lovasz-Pelikan index)[3] is a topological index that indicates the degree of branching in a molecule (the higher the value of LP1 the higher the number of ramification). The BELp1 descriptor (lowest eigenvalue n°1 of Burden matrix weighted by atomic polarizability)[21] belong to the BCUT descriptors family. BCUT descriptors were designed to encode atomic properties relevant to intermolecular interactions and have been widely used in QSAR models generation.[22,23] By using together, LP1 and BELp1 descriptors were expected to give simultaneous information about the shape (molecular branching) and the electronic properties (polarizability) of the molecules.

Employing the two descriptors, all the compounds in the MD and in the VSTS subsets have been therefore submitted to the *k*-means cluster analysis module of STATISTICA 6.0 software package.[24] This type of algorithm will start with a given *k* number of random clusters, and then move the objects (in this case the molecules) between those clusters with the goals to (1) minimize variability within clusters and (2) to maximize variability between clusters. Indeed, the *k*-means method will produce *k* different clusters of the greatest possible distinction. In fact, treating every single target individually for clustering would led to a large number of clusters (*n* clusters for 8 kinases) and consequently to the development of *n*×8 target-specific/cluster-specific QSAR models. Moreover, due to the *k*-means algorithm structure, the *n*th clusters of every kinases would not be comparable. Thus, in order to obtain multitarget models, the clusters have to be generated from the entire dataset. In this way, in fact, the clusterization process depends only on the inhibitors structures and not on both inhibitor structures and the considered kinases. The number of clusters have been selected in a manner that the number of cases for cluster fallen in a normal distribution (Figure 3). Moreover, in order to avoid the generation of too small subsets of compounds, we assumed that each cluster must contain at least 100 compounds. This goal was reached splitting the cases in six different subsets. In fact, the use of more clusters would lead to groups with a very few number of cases (even less than 20). On the contrary, the generation of 4 or 5 clusters led to a not normal distribution. As the regrouping process was not based on molecular scaffolds, it resulted also in a nonstructural clusterization of the data. Let us consider the molecules reported in Figure 4 as a clarifying example. On the basis of the herein proposed clusterization method, compounds **A** and **B** were considered as members of the same group (cluster 4), and thus were deemed to obey to same QSAR equation. On the contrary, compound **C** (that showed a high degree of structural similarity with compound **B**; see the Minimal Common Structure, MCS, depicted in Figure 4) was inserted in a different cluster. It is noteworthy that the commonly used similarity indices (e.g. Tanimoto's score or MCS) would have considered compounds **B** and **C** as member of the same group. Thus, compounds bearing the same scaffold were found to be member of several clusters, likewise each cluster contained compounds with different scaffolds (Table 2). The case distribution as function of clusters and kinases is reported in Table 3.

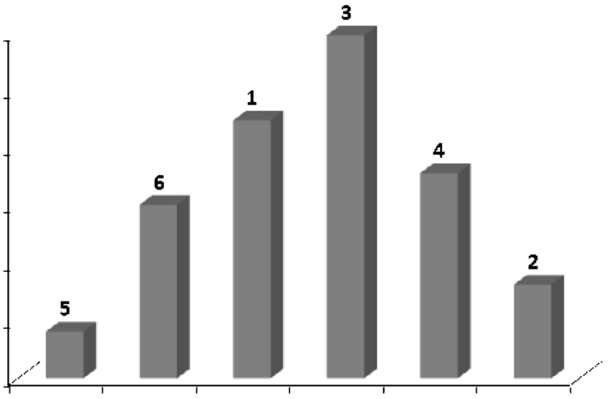

Figure 3. Number of cases for cluster. The cluster identification numbers are reported above the bars.

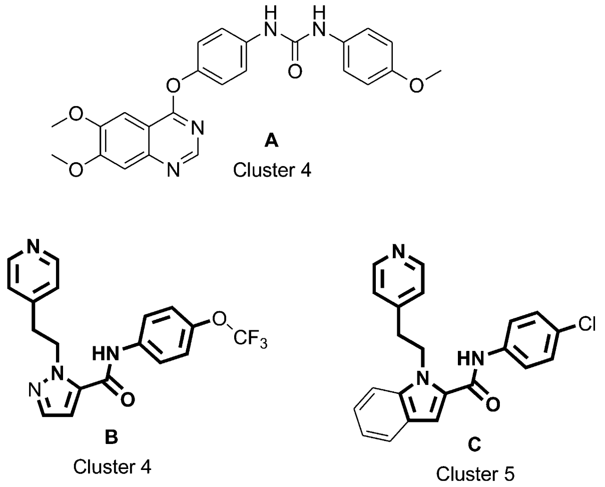

Figure 4. Cluster assignment based on BELp1 and LP1 values for three example compounds (A, B and C). For compounds B and C the Minimal Common Structure (MCS) have been highlighted with bold lines.

Several classification QSAR models have been derived starting from the MD. The molecular descriptors associated to the compounds have been retrieved from Dragon 5.5, which is a software widely used for QSAR studies. Moreover, since the models had to be also multi-target classifiers, several descriptors related to TKs were also assigned to each case. As Dragon was able to handle only molecules with at most 1000 heavy atoms (which correspond to a protein sequence of about 100 amino acids), the software was not employable for computing the enzyme related descriptors. The inconvenience has been resolved by use of an online server[25] that calculates the descriptors starting from the FASTA sequences of the proteins.

At this stage, we decided to edge the computation of regression models, i.e. models able to predict the pIC_{50} of the compounds. In fact, the IC_{50} was not exactly determined for all the considered compounds, since in a number of cases the activity was indicated generally as "higher than a value"

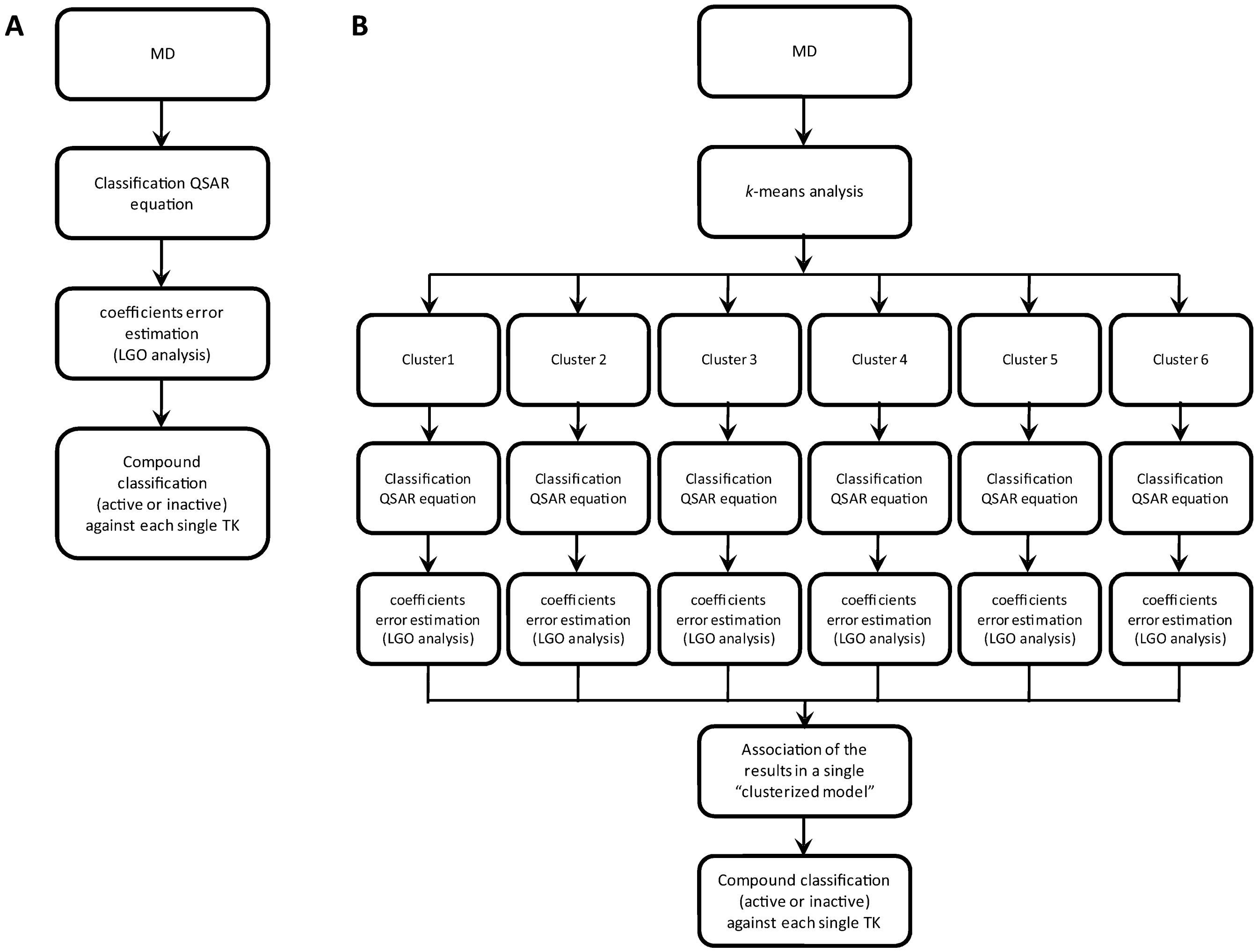
A
MD
Classification QSAR equation
coefficients error estimation (LGO analysis)
Compound classification (active or inactive) against each single TK
B
MD
k-means analysis
Cluster1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Classification QSAR equation
coefficients error estimation (LGO analysis)
Association of the results in a single "clusterized model"
Compound classification (active or inactive) against each single TK

Figure 5. A) Process employed for the development of the whole model classification QSAR. B) Strategy employed for the derivation of the clusterized classification QSAR model.

Table 2. Scaffolds vs. cluster compounds distribution.

Cluster ID	Scaffold ID[a]																									
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	1	–	1	26	–	41	–	–	–	15	10	–	–	5	7	111	22	1	–	30	–	2	23	10	–	10
2	20	–	20	–	–	2	–	1	–	–	–	–	–	10	1	–	4	1	–	–	–	35	–	–	–	–
3	–	88	–	–	12	75	25	–	57	–	1	59	–	–	–	5	2	51	–	–	48	–	1	–	–	–
4	40	–	–	–	–	207	2	–	1	–	–	–	–	–	–	–	5	7	36	–	–	1	–	–	–	–
5	7	–	–	–	–	–	–	–	–	1	1	–	40	–	–	1	1	–	–	–	–	1	–	1	–	–
6	1	1	–	–	1	16	–	15	–	27	5	–	93	–	–	–	1	–	–	–	–	–	–	–	14	–

[a] See Figure 2 for scaffolds representations.

(e.g. $IC_{50} > 10$ μM). The exclusion of these cases would leave in the dataset too few compounds assayed against more than one TK to allow the generation of statistically significant multi-target models. To bypass the problem, the compounds activity have been encoded by a binary dummy variable called Class (*C*), which allowed to classify the compounds as *active* or *inactive* depending on established activity thresholds. In this work, two cut-off of activity have been considered (1.0 μM and 0.1 μM respectively). These values of cut-off were chosen since the activity of a hit compound normally falls in this range (to find a hit is the major goal of a virtual screening process). For both the cut-

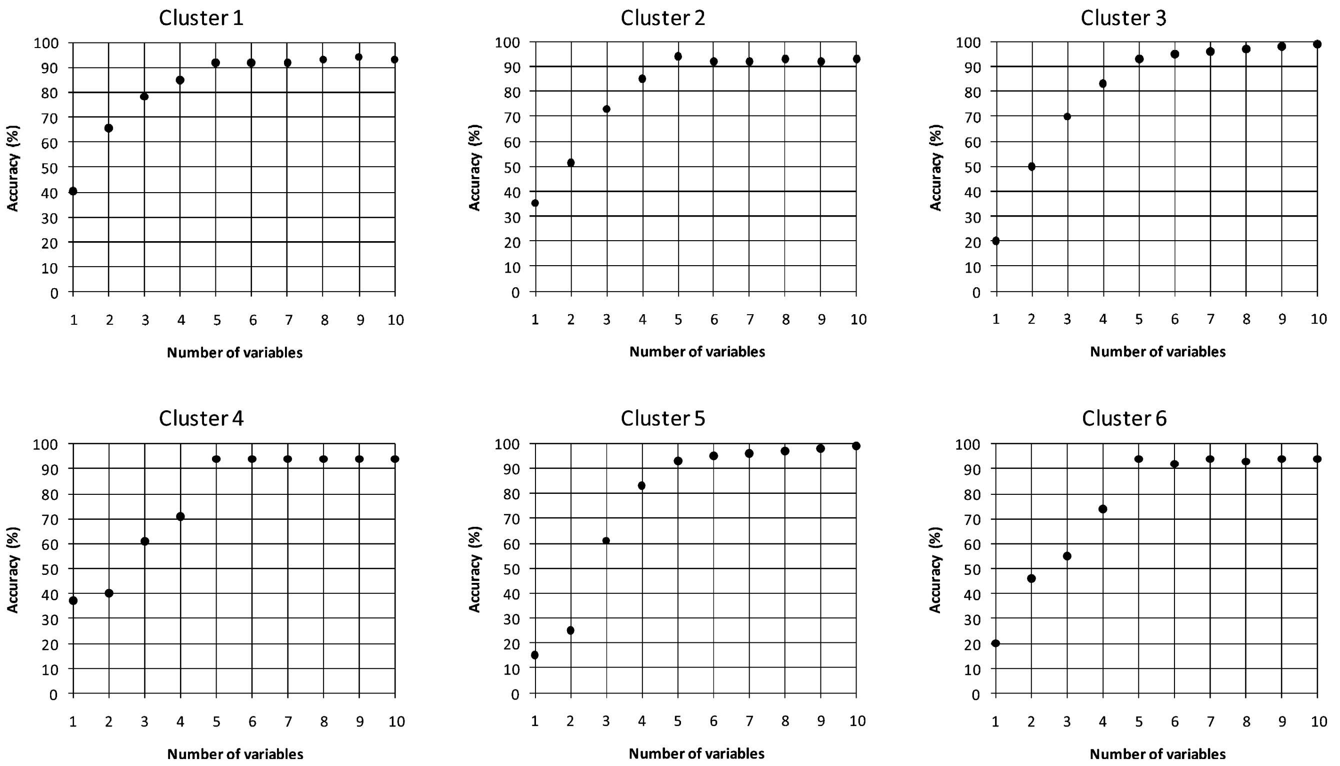

Figure 6. Dependence of the Accuracy on the number of independent variable used. For all the clusters, the use of more than 5 molecular descriptors did not lead to an appreciable increase in model Accuracy.

Table 3. TKIs distribution as function of TK and cluster.

Cluster ID	Number of inhibitors for each TK							
	abl	Lck	Src	EGFR	FGFR-1	KDR	PDGFRβ	VEGFR-1
1	58[a] + 0[b]	2 + 0	44 + 0	46 + 0	2 + 22	76 + 22	118 + 22	19 + 22
2	39 + 0	27 + 0	21 + 0	34 + 0	1 + 0	21 + 10	0 + 0	0 + 10
3	2 + 0	49 + 0	222 + 12	59 + 12	82 + 0	58 + 12	71 + 0	18 + 0
4	1 + 0	0 + 0	5 + 0	200 + 0	0 + 0	82 + 1	36 + 0	31 + 1
5	0 + 0	0 + 0	33 + 0	29 + 0	1 + 0	6 + 0	1 + 0	7 + 0
6	2 + 0	13 + 0	109 + 8	89 + 5	12 + 0	31 + 15	14 + 4	0 + 0

[a] The first number indicates the number of cases in the MD. [b] The second number indicates the number of cases in the VSTS.

off, a "whole model" and a "clusterized model" have been computed. The first ones were derived from a classical QSAR approach, i.e. the whole MD have been submitted to statistical analysis. Conversely, the "clusterized models" were gathered from the association of 6 cluster-specific multi-target QSARs (the processes for the QSAR equations computation are summarized in Figure 5).

As described in the Section 4, for the models computation, the General Discriminant Analysis (GDA) module of STATISTICA has been employed. The number of variables necessary for each model have been determined comparing the accuracies obtained with a number of descriptors ranging from 1 to 10 (see Figure 6). In all the cases, the use of more than 5 variables did not allow an appreciable enhancement in the models accuracy.

2.3 QSAR Models Structure and Usage

The QSAR model computed for cluster 1 with a cut-off of 0.1 μM is reported by means of example in Equation 1 (see Section 4.5 for a detailed description of the computation approach and the Supporting Information for all the equations and the molecular descriptor codes).

$$\begin{aligned}\text{Score} = {} & 4.57(\pm 0.53)\cdot \text{EEig07x}-1.77(\pm 0.39)\\ & \cdot \text{GGI7} + 2.51(\pm 0.32)\cdot \text{BLTD48} + 0.62(\pm 0.09)\\ & \cdot {}^{\text{ATP}}\text{pk1}_1 + 1.50(\pm 0.15)\cdot {}^{\text{ATP}}\text{pk2}_4-0.20(\pm 0.04)\end{aligned} \quad (1)$$

All the computed models contained both compound-related descriptors (EEig07x, GGI7 and BLTD48 in Equation 1) and TK-related descriptors (^{ATP}pk1$_1$ and ^{ATP}pk2$_4$ in Equation 1). Consequently, to predict the inhibitory profile of a compound, it is necessary: i) to determine to which cluster the compound belong to; ii) to select the appropriate QSAR equation, accordingly to the cluster number and the desired cut-off of activity; iii) to calculate 8 different score values, one for each kinase, inserting in the equation the appropriate molecular descriptors for both the compound and the kinase.

2.4 QSAR Models Evaluation

To test whether the models were useful in virtual screenings, they have been evaluated using the VSTS (Table 4 and 5). At both the cut-off of activity, a dramatic reduction in VSTS classification accuracy was found for the whole models (55.1% at 1.0 μM and 50.0% at 0.1 μM), which thus worked as random classifiers. On the contrary, the clusterization led to models that maintained an acceptable accuracy in discriminating between active and inactive compounds in the VSTS, even if to lower extent with respect to the MD. As depicted in Figure 7, the clusterization process led to an enhancement in the prediction accuracies for all the considered kinases with respect to the whole models. Moreover, the clusterized models showed, at both the cut-off of activity, an higher homogeneity in the accuracies referred to each single kinase. Consequently, the herein proposed strategy led to models able to predict in a more satisfactory manner the selectivity profile of the inhibitors. Also the single accuracies for each cluster were determined (Figure 8). In both the cases, the cluster 4 was predicted with the lower accuracy, while at a cut-off of 0.1 μM the cluster 6 showed a remarkable increase in accuracy. However this two clusters accounted respectively only for the 1.1% and the 18.0% of the cases in the VSTS. The remaining three clusters (that accounted for about the 81% of the cases in the VSTS) showed very similar accuracies. Taken together, these data demonstrate that the accuracy increase for the VSTS was due to an overall effect across all targets and all the clusters rather than to the better performance for one TK or for one cluster.

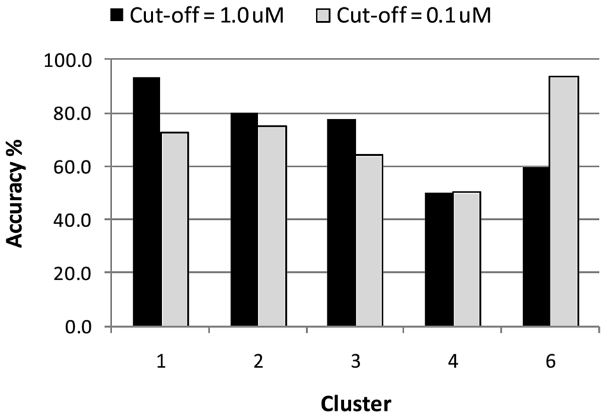

Figure 8. Accuracies for each cluster (VSTS prediction). Note that cluster 5 is not reported since no compound belong to this cluster in the VSTS.

Interestingly, although there were no compounds belonging to scaffold 2 assayed against VEGFR-1 in the MD (see Table 3), the QSAR models predicted in a very satisfactory manner those contained in the VSTS. In fact, the accuracies were about 80% and 70% at the cut-off of 1.0μM and 0.1 μM respectively. These values agreed with the mean accuracies computed for cluster 2 (see Figure 8). Thus, taking advantage of the multi-target approach (i.e. employing also TK-related descriptors) instead of computing TK-specific models, we generated models that were not affected by the lack of sample for a specific target in a cluster in the training set.

The higher performance of the clusterized model with cut-off of 0.1 μM to distinguish between active and inactive compounds in the VSTS has been also highlighted through the enrichment plot. As depicted in Figure 9, the whole models worked in a manner quite similar to a random classifier; on the contrary, the clusterized model worked in a satisfactory manner, as the result was much closer to the

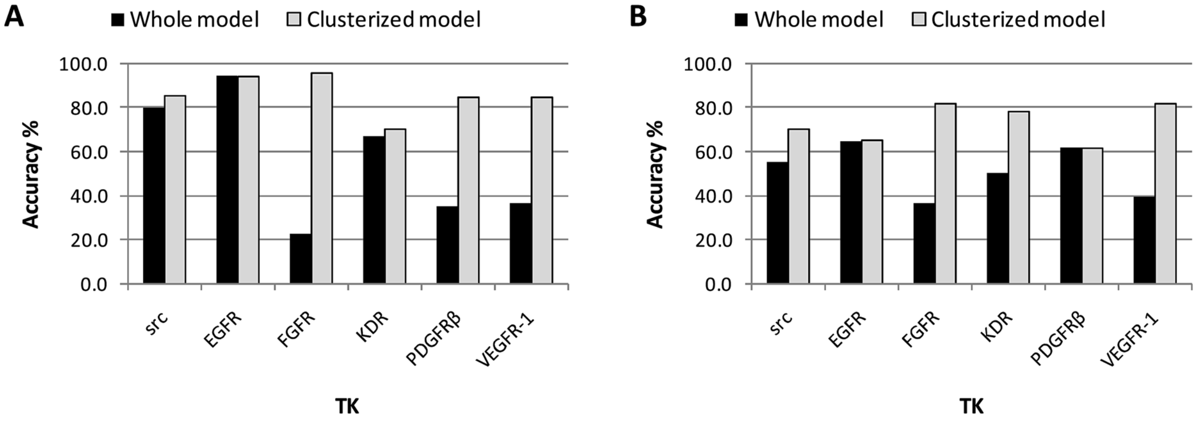

Figure 7. Comparison of the accuracies for each TK (VSTS prediction) between whole model and clusterized model. A: cut-off of 1.0 μM. B: cut-off of 0.1 μM. Note that abl and lck are not reported since no compound assayed against this TKs were present in the VSTS.

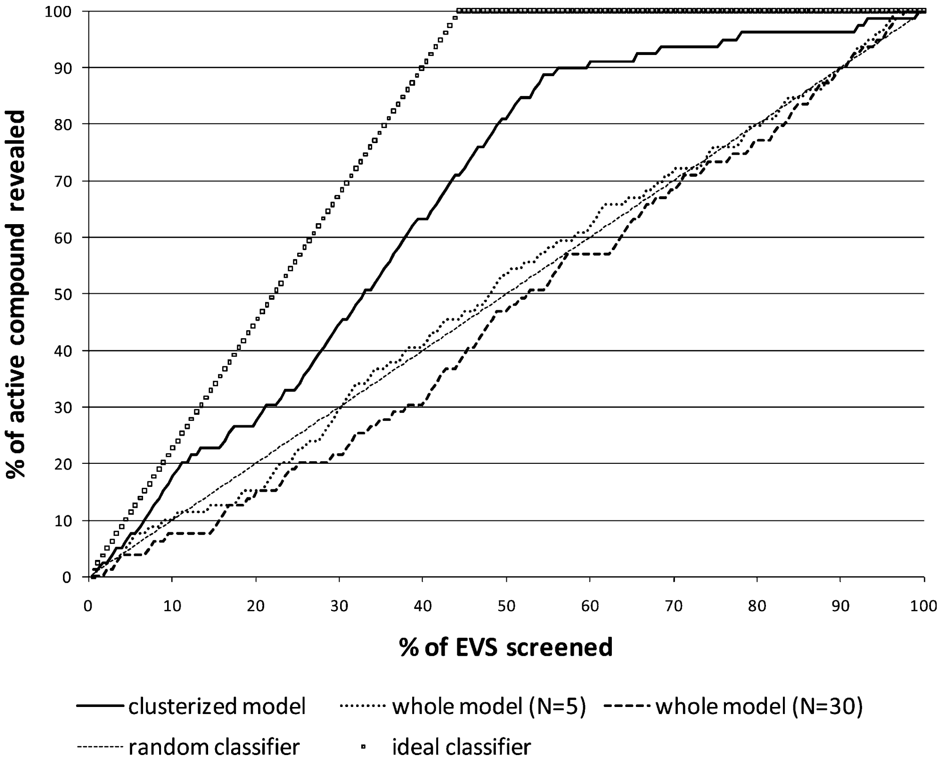

Figure 9. Enrichment plot (cut-off = 0.1 μM).

Table 4. QSAR models statistical parameters (cut-off = 1.0 μM).

	Whole model		Clusterized model[a]	
	MD	VSTS	MD	VSTS
Sensitivity (%)	83.5	32.6	91.0	69.2
Specificity (%)	84.6	80.7	91.1	85.6
Accuracy (%)	84.2	55.1	91.0	82.0
Number of false positive	120	16	66	20
Number of true positive	882	67	949	119
Number of false negative	160	64	93	12
Number of true negative	609	31	663	27
F-test *($p<0.05$)*	312.44		110.11	
Chi-squared	1016.73		261.87	
Wilk's lambda	0.48		0.35	

[a] Results arising from the six separated QSAR models have been joined in a single statistic for a more clear comparison with the not-clusterized model.

ideal classifier than to the random one. These data indicate that clusterized models could be employed for virtual screening purpose, while the same is not true for the whole models.

2.5 QSAR Models Validation

We also took into consideration the hypothesis that the improvement in performances could be ascribed to other two factors: the reduction in data dimensionality or the increase in the number of variables used. In fact, each cluster contained a fewer number of cases than the entire MD; moreover, the clusterized models were derived from the association of six independent models and thus a higher total number of variable was used. Indeed, to assess whether the enhancement in models accuracies was really due to the clusterization process, several other QSAR studies have been performed (at this stage only the cut-off of 0.1 μM was considered). First of all, we stochastically split the MD in six groups bearing the same cases number distribution than the clusters previously determined, and derived a QSAR models for each group. Recollection of the models

Table 5. QSAR models statistical parameters (cut-off = 0.1 μM).

	Whole model		Clusterized model[a]	
	MD	VSTS	MD	VSTS
Sensitivity (%)	82.7	53.9	87.5	89.7
Specificity (%)	82.6	39.6	88.6	65.5
Accuracy (%)	82.7	50.0	89.0	74.7
Number of false positive	176	29	127	38
Number of true positive	622	19	667	72
Number of false negative	131	60	86	7
Number of true negative	842	70	891	61
F-test *(p<*Mk*>0.05)*	248.25		120.35	
Chi-squared	876.50		365.91	
Wilk's lambda	0.52		0.45	

[a] Results arising from the six separated QSAR models have been joined in a single statistic for a more clear comparison with the not clusterized model.

results into a single "randomly clusterized model" led to accuracies of 84.3% in MD and 45.6% in VSTS. This values were significantly lower than the corresponding accuracies for the "clusterized model" (see Table 5). Then, a model based on the whole MD has been derived using 30 molecular descriptors, i.e. the total number of descriptors contained in the clusterized models. As well as observed in the previous experiment, the model showed good accuracy in training (91.5%), but it dramatically failed in classifying the compounds of the VSTS (47.8% of accuracy). Neither the reduced datasets dimensionality nor the increase in the number of molecular descriptors employed could be, then, considered responsible for the better performance of the clusterized models.

Finally, we tested our clusterized QSAR models through novel compounds classification (Figure 10). All the tested compounds had been previously synthesized by our research group. While the cytotoxic activities were determined for all the compounds,[26] nor the same was for the kinases selectivity profile. All the six derivatives shared the quinazoline scaffold and the clusterization process regrouped them in the 3rd cluster. All the compounds have been tested against EGFR, FGFR-1, KDR, PDGFRβ and abl. The model predictions, compared with those obtained with the whole models with 5 and 30 independent variables respectively, are reported in Table 6 and 7.

Also in this case, the clusterization process led to a model able to discriminate between active and inactive compounds in a more satisfactory manner with respect to the whole models. Moreover, it is noteworthy that while the whole models outputted a large number of false positive results, the same was not seen for the clusterized one.

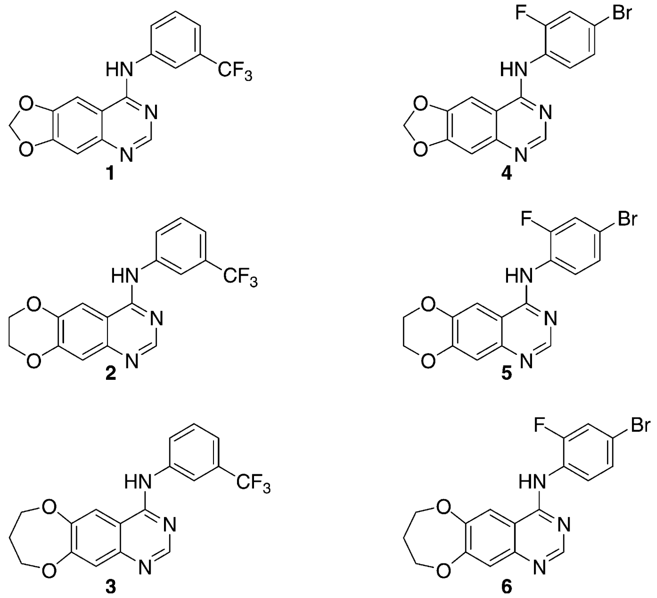

Figure 10. Compounds used in models evaluation.

Table 6. QSAR models evaluation for the six novel compounds (cut – off = 0.1 μM).

	Clusterized model	Whole model (*N* = 5) [a]	Whole model (*N* = 30) [a]
Sensitivity (%)	100.0	100.0	0.0
Specificity (%)	66.7	13.8	13.3
Accuracy (%)	93.3	16.7	13.3
Number of false positive	2	25	26
Number of false negative	0	0	0

[a] Number of independent variables used to derive the model.

3 Conclusions

In this work a novel approach for QSAR models computation has been developed. Through this novel strategy, several multi-target QSAR models for TKIs classification based on compounds clusterization have been developed, one for each cluster. The clusterization process hedged the use of similarity indexes or molecular fingerprints, as it was based

Table 7. mtc-QSAR models prediction for the six novel compounds (cut – off = 0.1 μM).

CPD	Predicted class[a]					Observed class[a]				
	EGFR	FGFR1	KDR	PDGFRβ	abl	EGFR	FGFR1	KDR	PDGFRβ	abl
1	0	0	0	0	0	0	0	0	0	0
2	0	0	1	1	0	0	0	0	1	0
3	0	0	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0
5	0	0	0	1	0	0	0	0	1	0
6	1	0	0	1	0	1	0	0	1	0

[a] 0 = compound observed or predicted as inactive ($IC_{50} > 0.1$ μM). 1 = compound observed or predicted as active ($IC_{50} < 0.1$ μM).

on molecular descriptors. In this way, the clusterization results was independent of structural features and thus it did not depend on the initial set of compounds used to generate the models. The importance of the clusterization has been demonstrated comparing the overall clusterized QSAR model prediction with those obtained from not-clusterized models characterized by different number of independent variables. The models have been compared using both an external validation set of compounds (VSTS) and novel derivatives. Because of the good accuracy of the model, the methodology could represent an useful tool to find novel promising kinases inhibitors. Moreover, since the models take into account a set of 8 different enzymes, it could be employed to design novel compounds with specific activity profile or in virtual screening processes. Finally, since the enhancement in prediction was demonstrated to be related only to the kind of clusterization process employed, the herein presented strategy could constitute a novel and interesting approach to review those models which failed in the prediction of activities of novel compounds or to derive new QSAR models useful for VS purpose. In fact, despite the increase in the computational cost, the clusterization process allowed the generation of models able to mine structurally heterogeneous datasets.

4 Experimental

4.1 Chemistry

Compounds **1–6** were synthesized as previously described.[26]

4.2 In Vitro Kinase Assays

Activated tyrosine kinases (abl, EGFR, FGFR-1, KDR and PDGFRβ) were purchased from Sigma Aldrich or Calbiochem and were diluted in kinase dilution buffer (5 mM MOPS pH 7.2, 2.5 mM glycerol 2-phosphate, 5 mM MgCl2, 0.4 mM EGTA, 0.4 mM EDTA, 0.05 mM DTT, 0.5 mM BSA). Reactions were set up in pre-cooled microcentrifuge tubes. Tested compounds have been added with: active kinase at final concentration of 200 ng/mL; 0.2 mg/ml substrate solution (Myelin Basic Protein, Sigma); 0.05 mM ATP and 0.25 μCi of [γ-32P]ATP (PerkinElmer, Monza, MI) to a final reaction volume of 25 μL. The final concentration of tested compounds was 0.1 μM. Negative controls were prepared replacing the substrate solution with water whereas positive controls were set up replacing test molecules with water. Reactions were carried out at 30 °C for 20 min and finally stopped by the addition of loading buffer containing 0.25 mM β-mercaptoethanol. Samples were then subjected to electrophoresis on 10% w/v SDS-PAGE gel. Gels were dried, and phosphorylated myelin basic protein was identified by autoradiography. The VersaDoc Quantity One software (BioRad) was used for densitometric analysis.

4.3 Datasets

The chemical structures of the known TKIs (1359 compounds) and their inhibitory activity (IC_{50} or pIC_{50}) were taken from literature (see Supporting Information for details). Only data regarding TKs cells free assays were taken in consideration and only if the assays were performed measuring the inhibition of a poly-Tyr synthetic peptide phosphorylation (i.e. no autophosphorylation assays were considered). Both reversible and irreversible TKIs were considered. The pIC_{50} values was converted to the corresponding IC_{50} (μM) values. The 2D structure of each compound was converted to the SMILES notation using MarvinSketch software.[27] The set of TKs includes 5 receptor TKs (EGFR, FGFR, KDR also named VEGFR-2, PDGFR and VEGFR-1) and 3 cytoplasmatic TKs (abl, src and lck).

Because of several compounds were tested against more than one TKs, the total amount of cases in the dataset was 1949. The compounds have been regrouped in two distinct datasets called the Main Dataset (MD) and the Virtual Screening Test Set (VSTS), using a cases split ratio of 10:1. The assignment of each case to the two subsets has been made in a manner that compounds belonging to the same reference work could be inserted only into the MD or the VSTS, i.e. there were no crossed references between the two subsets. By this way, the MD and the VSTS resulted constituted by compounds quite structurally unrelated. The MD contained 1771 cases (1300 compounds), while the VSTS contained the remaining 178 cases (59 compounds).

4.3.1 Tyrosine Kinases Primary Sequences

For each TK, the primary sequence of the entire protein and the one of the ATP-binding set were derived from the Pubmed database. The sequences were downloaded in FASTA format.

4.4 Molecular Descriptors

4.4.1 Molecular Descriptors for TKIs

The molecular descriptors for TKIs were calculated starting from the SMILES code employing Dragon 5.5 software. Only descriptors belonging to the 2D subtype were considered. A total amount of 527 descriptors were associated to each compound.

4.4.2 Molecular Descriptors for TKs

Starting from the FASTA sequences of both full protein and ATP-binding site, for each TK the Pseudo Amino Acid Compositions (PseAACs) were calculated through the on-line server of the Gordon life science institute.[26,28] The Type 1 PseAA mode has been employed, weighting the sequences through hydrophobicity-hydrophilicity, Mass, p*K*1, p*K*2 and p*I* indices. Weight factor and lambda parameter were set to 0.05 and to 10 respectively.

4.5 Cluster Analysis and QSAR Modeling

All the compounds were clusterized employing the *k*-means clustering module of STATISTICA 6.0 software. The parameters have been set as follow: number of cluster = 6; number of iterations = 10. The initial cluster centers have been determined through the "*sort distances and take observations at constant intervals*" option.

Each classification QSAR model was derived starting from the MD, employing the general discriminant analysis (GDA) module of STATISTICA 6.0 software. To each case, a dummy dependent variable (Class, *C*) was assessed, being $C=1$ when the compound showed an IC_{50} lower than the established cut-off of activity (*active compound*), and $C=0$ when the compound showed an IC_{50} higher than the cut-off (*inactive compound*). The discriminant function analysis implemented in GDA can be used both (1) to select those variables which mostly allow to discriminate between two groups and (2) to compute (with the selected variables) an equation able to predict to which group a compound belong. The methods, thus, furnish a simple and useful classification equation.

The subset variable selection was performed employing the *forward stepwise* algorithm. The forward stepwise algorithm is an iterative process which extracts exactly one variable at each computational step. Indeed, the number of selected variables is equal to the number of steps performed. The number of variables for models was set to 5 (see Section 2 for explanation). Once the selection of the variables has been performed, the GDA outputs two linear combinations of the descriptors, which allow to compute two classification scores for each compound:

$$S^0 = a^0 \cdot \mu_1 + b^0 \cdot \mu_2 + \ldots m^0 \cdot \mu_m + k^0 \qquad (2)$$

$$S^1 = a^1 \cdot \mu_1 + b^1 \cdot \mu_2 + \ldots m^1 \cdot \mu_m + k^1 \qquad (3)$$

where: S^0 and S^1 indicate the scores, $\mu_{1,2\ldots m}$ indicate the molecular descriptors, and 0 and 1 indicate to which class belong the terms.

Each compound should be classified as belonging to the group for which it showed the highest classification score. However, to give more readily usable classification functions, we simply combined the Equations 2 and 3 for each model:

$$S = S^1 - S^0 = (a^1 - a^0) \cdot \mu_1 + (b^1 - b^0) \cdot \mu_2 + \ldots + k^1 - k^0 \qquad (4)$$

As a consequence, those compounds whose score was higher than 0 have been classified as active, while those compounds whose score was lower than 0 resulted as inactive.

Following a procedure quite similar to that previously reported by Maw et al.[29] after a model has been computed, the relative dataset has been randomly split in a *training set* and a *validation set*, employing a 3:1 split ratio. Using the five variables previously determined, new coefficients were obtained by GDA on the *training set*. Then, the score values for the compounds in the *internal test set* were predicted. The overall process was repeated for ten times. The final score for each compound was determined as the average values obtained after this LGO process. Moreover, also the final equation coefficients were determined as the average values (±standard deviation) obtained after the LGO process.

Acknowledgements

The present work has been carried out with the financial support of the *University of Padova* "Progetto di Ricerca di Ateneo 2008" (CPDA084954). Marzaro, G. thanks financial support from the *University of Padova* for a postdoc grant.

References

[1] C. Hansch, T. Fujita, *J. Am. Chem. Soc.* **1964**, *86*, 1616.

[2] J. C. Dearden, *Expert Opin. Drug Metab. Toxicol.* **2007**, *3*, 635.

[3] R. Todeschini, *Handbook of molecular descriptors*, Wiley-WCH, **2002**.

[4] K. Hasegawa, Y. Miyashita, K. Funatsu, *J. Chem .Inf. Comput. Sci.* **1997**, *37*, 306.

[5] H. Wold, Multivariate Analysis (Ed: P. R. Krishnaiah), New York, Academic Press, **1966**, pp. 391.

[6] R. Burbidge, M. Trotter, B. Buxton, S. Holden, *Comput. Chem.* **2001**, *26*, 5.

[7] H. Yuan, A. L. Parrill, *Bioorg. Med. Chem.* **2002**, *10*, 4169.

[8] C. L. Senese, A. J. Hopfinger, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2180.
[9] J. Mendelsohn, *J. Clin. Oncol.* **2002**, *20*, 1S.
[10] J. Dowell, J. D. Minna, P. Kirkpatrick, *Nat. Rev. Drug Discov.* **2005**, *4*, 13.
[11] J. R. Johnson, P. Bross, M. Cohen, M. Rothmann, G. Chen, A. Zajicek, J. Gobburu, A. Rahman, A. Staten, R. Pazdur, *Clin. Cancer Res.* **2003**, *9*, 1972.
[12] S. Knapper, A. K. Burnett, T. Littlewood, W. J. Kell, S. Agrawal, R. Chopra, R. Clark, M. J. Levis, D. Small, *Blood* **2006**, *108*, 3262.
[13] E. Jabbour, J. Cortes, H. Kantarjian, *Expert. Opin. Investig. Drugs* **2007**, *16*, 679.
[14] M. Z. Dieter, S. L. Freshwater, W. A. Solis, D. W. Nebert, T. P. Dalton, *Biochem. Pharmacol.* **2001**, *61*, 215.
[15] H. Izzedine, I. Buhaescu, O. Rixe, G. Deray, *Cancer Chemother. Pharmacol.* **2007**, *60*, 357.
[16] M. Krug, A. Hilgeroth, *Mini Rev. Med. Chem.* **2008**, *8*, 1312.
[17] A. Petrelli, S. Giordano, *Curr. Med. Chem.* **2008**, *15*, 422.
[18] G. Marzaro, A. Chilin, A. Guiotto, E. Uriarte, P. Brun, I. Castagliuolo, F. Tonus, H. Gonzàlez-Diaz, *Eur. J. Med. Chem.* **2011**, *46*, 2185
[19] See Supporting Information for structure, biological activities and bibliographic references
[20] *DRAGON for Windows*, Version 5.5, Talete srl, **2007**, available from: http://www.talete.mi.it/.
[21] A. R. Leach, *An Introduction to Chemoinformatics*, Springer, Heidelberg **2007**.
[22] H. Gao, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 402.
[23] M. P. Gonzalez, C. Teran, M. Teijeira, P. Besada, M. J. Gonzalez-Moa, *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3491.
[24] *Statistica*, Version 6.0, StatSoftInc, 2002
[25] *PseAAC calculation*, available at http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/
[26] A. Chilin, M. T. Conconi, G. Marzaro, A. Guiotto, L. Urbani, F. Tonus, P. Parnigotto, *J. Med. Chem.* **2010**, *53*, 1862.
[27] *CHEMAXON MARVIN,* Version 5.1.3.2, Budapest, Software available at www.chemaxon.com/products.htm
[28] H. B. Shen, K. C. Chou, *Anal. Biochem.* **2008**, *373*, 386.
[29] H. H. Maw, L. H. Hall, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1270.

Received: March 8, 2011
Accepted: June 27, 2011
Published online: August 2, 2011