

A Novel QSAR Model for Evaluating and Predicting the Inhibition Activity of Dipeptidyl Aspartyl Fluoromethylketones

Antreas Afantitis^{a,b}, Georgia Melagraki^a, Haralambos Sarimveis^{a*}, Panayiotis A. Koutentis^c, John Markopoulos^d and Olga Igglessi-Markopoulou^a

^a School of Chemical Engineering, National Technical University of Athens, Athens, Greece

^b Department of ChemoInformatics, NovaMechanics Ltd., Cyprus

^c Department of Chemistry, University of Cyprus, P. O. Box 20537, 1678 Nicosia, Cyprus

^d Department of Chemistry, University of Athens, Athens, Greece

Keywords: Apoptosis, Caspase-3, Molecular Modeling, QSAR

Received: December 6, 2005; Accepted: April 11, 2006

DOI: 10.1002/qsar.200530208

Abstract

A linear quantitative structure activity relationship model is obtained using Multiple Linear Regression (MLR) analysis as applied to a series of 49 dipeptidyl aspartyl fluoromethylketone derivatives with inhibitory activity of the caspase enzyme. For the selection of the best descriptors, the elimination selection stepwise regression method is utilized. The accuracy of the proposed MLR model is illustrated using the following evaluation techniques: cross validation, validation through an external test set, and Y-randomization. Furthermore, the domain of applicability which indicates the area of reliable predictions is defined.

1 Introduction

Novel medicines are typically developed using a trial-and-error approach which is costly and time-consuming. The application of Quantitative–Structure Activity Relationship (QSAR) methodologies to this problem has the potential to greatly decrease the time and effort required to improve current medicines in terms of their efficacy or to discover new ones. QSAR constitutes an attempt to reduce the trial-and-error element in the design of compounds, by establishing mathematical relationships between physical, chemical, biological, or environmental activities of interest and measurable or computable parameters such as topological, physicochemical, stereochemistry, or electronic indices [1–6].

Apoptosis is the vital process by which cells undergo “programed cell death” in various biological systems. Diverse groups of molecules are involved in the apoptosis pathway. One set of mediators implicated in apoptosis belongs to the aspartate-specific cysteinyl proteases or caspases [7–9]. Caspases are a family of proteases that relay a “doomsday” signal in a stepwise manner reminiscent of signaling by kinases. Caspases are present in all cells as latent enzymes. A member of this family, caspase-3 has been identified as being a key mediator of apoptosis of mammalian cells [10]. Excessive apoptosis is responsible, at least in part, for a variety of diseases for example liver disease

[11], brain ischemia [12], myocardial infraction [13], Huntington’s disease, and Alzheimer’s disease [14].

In the past, two attempts have been made to build QSAR models in the general field of apoptosis. Hansch *et al.* [15] presented a QSAR study containing a variety of phenolic compounds causing apoptosis and later [16] the same scientific group presented a QSAR of apoptosis induction in various cancer cells.

In this study we utilized 49 dipeptidyl aspartyl fluoromethylketones [17–19] aiming at the investigation of their role as inhibitors of caspase-3 enzyme and the development of a QSAR model. Sixty-one physicochemical and topological descriptors were considered as input candidates to the model. The descriptors were calculated using Topix (www.lohninger.com/topix.html) and ChemSar which is included in the ChemOffice (CambridgeSoft Corporation) suite of programs. A rigorous variable selection procedure was adopted to define a small set of statistically significant physicochemical and topological descriptors that can determinate and predict the activity of the compounds that consisted our dataset. The QSAR models were obtained by Multiple Linear Regressions (MLRs). The result of this study is the development of a new linear QSAR model containing four variables. In order to validate the proposed methodology, we used two validation strategies: Y-randomization and external validation using division of the entire dataset into training and test sets.

2 Materials and Methods

2.1 Dataset

In this QSAR study 49 biological data from [17–19] were used. In order to model and predict the specific activity (inhibition of caspase-3), 61 physicochemical constants, topological, and structural descriptors (Table 1) were considered as possible input candidates to the model. All the descriptors were calculated using Topix and ChemSar.

The objective of this work was to determine the best variables which afford the most significant linear QSAR models linking the structure of these compounds with their inhibitory activity.

2.2 Stepwise Multiple Regression

As mentioned in Section 1, the Elimination Selection Stepwise Regression (ES-SWR) algorithm [20] was used to select the most appropriate descriptors. ES-SWR is a popular stepwise technique that combines Forward Selection (FS-SWR) and Backward Elimination (BE-SWR). It is basically a forward selection approach, but at each step it considers the possibility of deleting a variable as in the

backward elimination approach, provided that the number of model variables is greater than 2. The two basic elements of the ES-SWR method are described next in more detail.

2.2.1 Forward Selection

According to the standard forward selection algorithm, the variable considered for inclusion at any step is the one yielding the largest single degree of freedom F -ratio among the variables that are eligible for inclusion. The variable is included only if the corresponding F -ratio is larger than a fixed value F_{in} . Consequently, at each step, the j th variable is added to a k -size model if

$$F_j = \max_j \left(\frac{RSS_k - RSS_{k+j}}{s_{k+j}^2} \right) > F_{in} \quad (1)$$

In the above inequality RSS is the Residual Sum of Squares and s is the mean square error. The subscript $k+j$ refers to quantities computed when the j th variable is added to the k variables that are already included in the model.

Table 1. Physicochemical constants, topological and structural descriptors.

ID	Description	Notation	ID	Description	Notation
1	Molar Refractivity	MR	2	Diameter	Diam
3	Partition Coefficient (Octanol Water)	Clog P	4	Molecular Topological Index	TIndx
5	Principal Moment of Inertia Z	PMIZ	6	Number of Rotatable Bonds	NRBo
7	Principal Moment of Inertia Y	PMIY	8	Polar Surface Area	PSAr
9	Principal Moment of Inertia X	PMIX	10	Radius	Rad
11	Connolly Accessible Area	SAS	12	Shape attribute	ShpA
13	Connolly Molecular Area	MS	14	Shape coefficient	ShpC
15	Total Energy	TotE	16	Sum of Valence Degrees	SVDe
17	LUMO Energy	LUMO	18	Total Connectivity	TCon
19	HOMO Energy	HOMO	20	Total Valence Connectivity	TVCon
21	Balaban Index	BIndx	22	Wiener Index	WIndx
23	Cluster Count	ClsC	24	Randic 0	Chi0
25	Randic 1	Chi1	26	Randic 2	Chi2
27	Randic 3	Chi3	28	Randic 4	Chi4
29	Randic Information 0	ChiInf0	30	Randic Information 1	ChiInf1
31	Randic Information 2	ChiInf2	32	Randic Information 3	ChiInf3
33	Randic Information 4	ChiInf4	34	Kier-Hall 0	Ki0
35	Randic Mod	ChiMod	36	Xu1	Xu1
37	Xu2	Xu2	38	Xu3	Xu3
39	Balaban Topological	TopoJ	40	Topological Radius	TopoRad
41	Topological Diameter	TopoDia	42	Number of Branches	NBranch
43	Number of Rings	NRings	44	Wiener Dim	Wiener Dim
45	Bertz	Bertz	46	AtomCompMean	\bar{I}_{AC}
47	AtomCompTot	AtomCompTot	48	Zagreb1	Zagreb1
49	Zagreb2	Zagreb2	50	Quadratic	Quadr
51	ScHultz	ScHultz	52	Kappa1	Kappa1
53	Kappa3	Kappa3	54	Kappa2	Kappa2
55	Wiener Distance	WienerDistCode	56	Wiener Information	InfWiener
57	DistEqMean	DistEqMean	58	DistEqTotal	DistEqTotal
59	InfMagnitDistTot	InfMagnitDistTot	60	Polarity	Polarity
61	Gordon	Gordon			

We have slightly modified the above algorithm in order to ensure that the selected variables are not highly inter-correlated. More specifically, a variable is added to a k -size model if the criterion described by Eq. 1 is satisfied and additionally all the correlation coefficients with the k variables that have already been selected by the algorithm are below a fixed value.

2.2.2 Backward Elimination

The variable considered for elimination at any step is the one yielding the minimum single degree of freedom F -ratio among the variables that are included in the model. The variable is eliminated only if the corresponding F -ratio does not exceed a specified value F_{out} . Consequently, at each step, the j th variable is eliminated from the k -size model if

$$F_j = \min_j \left(\frac{\text{RSS}_{k-j} - \text{RSS}_k}{s_k^2} \right) < F_{\text{out}} \quad (2)$$

The subscript $k-j$ refers to quantities computed when the j th variable is eliminated from the k variables that have been included in the model so far.

2.3 Y-Randomization Test

This technique ensures the robustness of a QSAR model [21, 22]. The dependent variable vector (biological action) is randomly shuffled and a new QSAR model is developed, including the selection of the best possible variables using the ES-SWR algorithm. The procedure is repeated several times and the new QSAR models are expected to have low R^2 and R_{CV}^2 values. If the opposite happens then an acceptable QSAR model cannot be obtained for the specific modeling method and data.

2.4 Estimation of the Predictive Ability of a QSAR Model

According to Tropsha group [22, 25, 26] a QSAR model is considered predictive, if the following conditions are satisfied:

$$R^2 > 0.6 \quad (3)$$

$$\frac{(R^2 - R_0^2)}{R^2} < 0.1 \text{ or } \frac{(R^2 - R_0'^2)}{R^2} < 0.1 \quad (4)$$

$$0.85 \leq k \leq 1.15 \text{ or } 0.85 \leq k' \leq 1.15 \quad (5)$$

Mathematical definitions of R_0^2 , $R_0'^2$, k , and k' are based on regression of the observed activities against predicted activities and the opposite (regression of the predicted activities against observed activities). The definitions are presented clearly in [22] and are not repeated here for brevity.

2.5 Defining Model Applicability Domain

In order for a QSAR model to be used for screening new compounds, its domain of application [23, 24] must be defined and predictions for only those compounds that fall into this domain may be considered reliable. *Extent of extrapolation* [22] is one simple approach to define the applicability of the domain. It is based on the calculation of the leverage h_i [25] for each chemical, where the QSAR model is used to predict its activity:

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (6)$$

In Eq. 6, x_i is the descriptor-row vector of the query compound and X is the $k \times n$ matrix containing the k descriptor values for each one of the n training compounds. A leverage value greater than $3k/n$ is considered large. It means that the predicted response is the result of a substantial extrapolation of the model and may be not reliable.

3 Results and Discussion

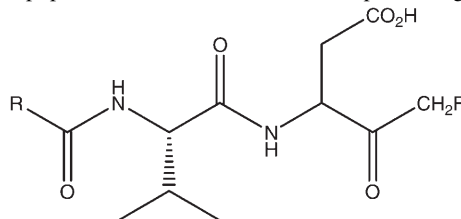
For the selection of the most important descriptors, the aforementioned stepwise multiple regression technique was used. In order to automate the procedure, we have developed in-house software that realizes the modified ES-SWR algorithm. The most significant descriptor according to the ES-SWR algorithm is the lipophilicity (Clog P) followed by Highest-Occupied Molecular Orbital (HOMO) and Lowest-Unoccupied Molecular Orbital (LUMO) energies and mean information index on atomic composition. The four above-mentioned descriptors are not highly correlated (Table 2).



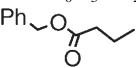
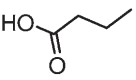
All the structures before the calculation of the descriptors were fully optimized using CS mechanics and more specifically MM2 force fields and the Truncated-Newton – Raphson optimizer, which provide a balance between speed and accuracy [20].

Lipophilicity is known to be important for absorption, permeability, and *in vivo* distribution of organic compounds [26, 27] and has been used as a physicochemical descriptor in QSARs with great success [28, 29]. Molecular Orbital (MO) surfaces visually represent the various stable electron distributions of a molecule. According to Frontier orbital theory, the shapes and symmetries of the HOMO

Table 2. Correlation matrix of the four selected descriptors.

	Clog P	HOMO	LUMO	ACM
Clog P	1.00			
HOMO	0.41	1.00		
LUMO	-0.32	0.12	1.00	
ACM	-0.04	0.34	-0.01	1.00

Table 3. Caspase-3 inhibiting activity of the dipeptide inhibitors. SAR of the *N*-protecting group.

	R	IC ₅₀ (nM) (exp)	log(1/IC ₅₀) (exp)	Leverages (limit 0.31)	log(1/IC ₅₀) (calc. by Eq. 8)	Training data log(1/IC ₅₀) (calc. by Eq. 9)	Validation data log(1/IC ₅₀) (calc. by Eq. 9)
1	CH ₃	250	-2.40	0.15	-2.00	-2.10	
2	CH ₃ CH ₂	81	-1.91	0.11	-1.95	-2.03	
3*	PhCH ₂	61	-1.78	0.09	-1.99		-2.06
4	CH ₃ O	37	-1.57	0.15	-1.85	-1.93	
5	PhCH ₂ CH ₂	98	-1.99	0.11	-1.88	-1.96	
6		35	-1.54	0.12	-1.55	-1.64	
7*		30	-1.48	0.08	-1.84		-1.89
8	PhCH ₂ CH ₂ O	110	-2.04	0.09	-1.59	-1.67	
9*	PhCH ₂ CH ₂ CH ₂ O	46	-1.66	0.12	-1.50		-1.59
10*	2-Cl-C ₆ H ₄ CH ₂ O	36	-1.56	0.05	-1.32		-1.38
11	3-Cl-C ₆ H ₄ CH ₂ O	36	-1.56	0.05	-1.35	-1.41	
12*	4-Cl-C ₆ H ₄ CH ₂ O	34	-1.53	0.05	-1.42		-1.47
13	2-F-C ₆ H ₄ CH ₂ O	38	-1.58	0.02	-1.65	-1.71	
14	3-F-C ₆ H ₄ CH ₂ O	29	-1.46	0.02	-1.69	-1.74	
15*	4-F-C ₆ H ₄ CH ₂ O	28	-1.45	0.02	-1.64		-1.69
16	2,4-di-Cl-C ₆ H ₃ CH ₂ O	25	-1.40	0.15	-1.23	-1.27	
17	3,4-di-Cl-C ₆ H ₃ CH ₂ O	21	-1.32	0.13	-1.27	-1.31	
18*	2,5-di-Cl-C ₆ H ₃ CH ₂ O	15	-1.18	0.13	-1.17		-1.21
19	2,4-di-F-C ₆ H ₃ CH ₂ O	35	-1.54	0.05	-1.78	-1.80	
20*	3,4-di-F-C ₆ H ₃ CH ₂ O	30	-1.48	0.05	-1.83		-1.84
21		33	-1.54	0.08	-1.58	-1.66	
22*		50	-1.70	0.13	-1.93		-2.01
23*	PhCH ₂ O	30	-1.48	0.05	-1.56		-1.64

and LUMO energies are crucial in predicting the reactivity of a species and the stereochemical and regiochemical outcome of a chemical reaction [20]. Before calculating the HOMO and LUMO energies (eV) all the structures were additionally fully optimized using the AM1 basis set. Mean information content on atomic composition \bar{I}_{AC} [20] is the mean value of the total information content and is calculated as

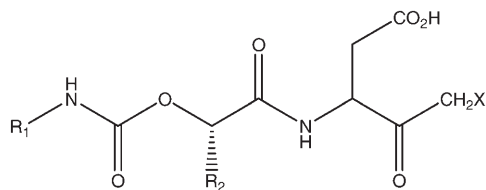
$$\bar{I}_{AC} = - \sum_g \frac{A_g}{A^h} \log_2 \frac{A_g}{A^h} = - \sum_g p_g \log_2 p_g \quad (7)$$

where A^h is the total number of atoms (hydrogen included), A_g is the number of equal-type atoms in the g th equiv-

alence class, and p_g is the probability of randomly selecting a g th type atom.

In order to investigate the possible existence of outliers, the extent of the extrapolation method was applied to the 49 compounds that constitute the entire dataset (Tables 3–5). The leverages for all the 49 compounds were computed (Tables 3–5) and one compound (id 36) was found to lie outside the domain of the model (warning leverage limit 0.31). This is justified by noticing that this specific compound (id 36) has a sufficiently more complex substituent in place of the fluoromethylketone. The compound was excluded from the rest of the analysis.

The full linear equation for the prediction of the inhibitory IC₅₀ activity is the following:

Table 4. Caspase-3 inhibiting activity of the dipeptide inhibitors. Peptidomimetic replacement of the P₂ α-amino acid.

	R ₁	R ₂	X	IC ₅₀ (nM) (exp)	log(1/IC ₅₀) (exp)	Leverages (limit 0.31)	log(1/IC ₅₀) (calc. by Eq. 8)	Training data log(1/IC ₅₀) (calc. by Eq. 9)	Validation data log(1/IC ₅₀) (calc. by Eq. 9)
24	Ph	Me	F	66	-1.82	0.17	-1.72	-1.78	
25*	Ph	2-Pr	F	17	-1.23	0.12	-1.44		-1.52
26	Ph	Cyclohexyl	F	50	-1.70	0.15	-1.28	-1.35	
27	PhCH ₂	Me	F	70	-1.85	0.04	-1.93	-1.98	
28*	PhCH ₂	2-Pr	F	20	-1.30	0.04	-1.67		-1.74
29	3-F-C ₆ H ₄	2-Pr	F	6	-0.78	0.06	-1.20	-1.28	
30	4-F-C ₆ H ₄	2-Pr	F	19	-1.28	0.10	-1.15	-1.24	
31*	3,4-diF-C ₆ H ₃	2-Pr	F	14	-1.15	0.10	-1.31		-1.36
32	2,4-diCl-C ₆ H ₃	2-Pr	F	14	-1.15	0.16	-0.61	-0.70	
33*	2,5-diCl-C ₆ H ₃	2-Pr	F	5	-0.70	0.15	-0.77		-0.86
34	4-PhO-C ₆ H ₄	2-Pr	F	12	-1.08	0.24	-1.06	-1.14	
35	Ph	2-Pr	DCB ^b	20	-1.30	0.24	-1.78	-1.74	
36 ^a	Ph	2-Pr	PTP ^c	70	-1.85	0.41	-	-	

^a Rejected from the dataset as outlier.

^b 2,6-dichlorobenzoyloxy.

^c 1-phenyl-3-(trifluoromethyl)pyrazol-5-yloxy.

$$\log(1/IC_{50}) = -2.78 + 0.32 \text{ Clog } P + 0.32 \text{ HOMO} + 1.16 \text{ LUMO} + 2.67 \bar{I}_{AC} \quad (8)$$

$$n = 48, s = 0.30, R^2 = 0.78, F = 38.65, R_{CV}^2 = 0.73, S_{PRESS} = 0.34$$

In order to further explore the prediction ability of the selected descriptors, the dataset of 48 dipeptidyl aspartyl fluoromethylketone derivatives was divided into a training set of 31 compounds and a validation set of 17 compounds. The selection of the derivatives in the training set was made according to the Kennard and Stone [30, 31] algorithm. The Kennard and Stone algorithm has gained an increasing popularity for splitting datasets into two subsets. The algorithm starts by finding two samples that are the farthest apart from each other on the basis of the input variables in terms of some metric, *e.g.*, the Euclidean distance. These two samples are removed from the original dataset and put into the calibration dataset. This procedure is repeated until the desired number of samples has been reached in the calibration set. The advantages of this algorithm are that the calibration samples map the measured region of the input variable space completely with respect to the induced metric and that all the test samples fall inside the measured region. According to Golbraikh and Tropsha [23] and Wu *et al.* [31], Kennard and Stone algorithm is one of the best ways to build training and test sets.

The compounds that constituted the training and validation sets are presented in Tables 3–5. The validation examples are marked with an asterisk. The rest of the study will be concentrated on the model which is constructed from the training set and will examine the predictive ability of the produced model. Using the four above-mentioned descriptors, we developed a new MLR equation based on only the 31 training examples:

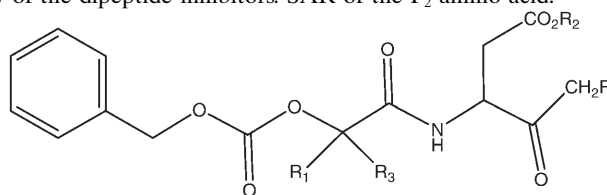
$$\log(1/IC_{50}) = -3.21 + 0.33 \text{ Clog } P + 0.26 \text{ HOMO} + 1.11 \text{ LUMO} + 2.52 \bar{I}_{AC} \quad (9)$$




$$n = 31, s = 0.33, R^2 = 0.75, F = 20.60, R_{CV}^2 = 0.65, S_{PRESS} = 0.39$$

Eq. 9 was used to predict the inhibitory activity of the caspase enzyme for the validation examples. The results are presented in Figure 1 and in the last column of Tables 3–5, which corresponds to $R_{pred}^2 = 0.86$. The results illustrated once more that the linear MLR technique combined with a successful variable selection procedure is adequate to generate an efficient QSAR model for predicting the inhibitory activity of different compounds.

The proposed model (Eq. 9) passed all the tests for the predictive ability (Eqs. 3–5)

$$R^2 = 0.86 > 0.6$$

Table 5. Caspase-3 inhibiting activity of the dipeptide inhibitors. SAR of the P₂ amino acid.

	R ₁	R ₂	R ₃	IC ₅₀ (nM) (exp)	log(1/IC ₅₀) (exp)	Leverages (limit 0.31)	log(1/IC ₅₀) (calc. by Eq. 8)	Training data log(1/IC ₅₀) (calc. by Eq. 9)	Validation data log(1/IC ₅₀) (calc. by Eq. 9)
37	i-Pr	Me	–	1100	–3.04	0.08	–2.39	–2.37	
38	i-Bu	H	–	200	–2.30	0.08	–2.40	–2.37	
39*	Bn	H	–	400	–2.60	0.10	–2.51		–2.49
40*	Me	H	–	600	–2.78	0.07	–2.69		–2.66
41*	H	H	–	1900	–3.28	0.09	–2.76		–2.72
42		H	–	1600	–3.20	0.26	–3.36	–3.31	
43		H	–	1400	–3.15	0.13	–2.99	–2.96	
44	Et	H	–	100	–2	0.06	–2.57	–2.54	
45	Ph	H	–	100	–2	0.09	–2.62	–2.59	
46	cyclohexyl	H	–	100	–2	0.14	–2.29	–2.27	
47		H	–	150	–2.18	0.05	–2.04	–2.04	
48	Me	H	Me	1400	–3.15	0.07	–2.64	–2.62	
49	i-Pr	H	Me	2300	–3.36	0.09	–2.67	–2.63	

$$\frac{(R^2 - R_0^2)}{R^2} = -0.28 < 0.1 \text{ or } \frac{(R^2 - R_0'^2)}{R^2} = -0.34 < 0.1$$

$$k = 0.96 \text{ and } k' = 1.02$$

For a more exhaustive testing of the predictive power of the model, except for the classical LOO cross validation technique, the validation of the model was carried out by a leave five out cross-(L5O) validation procedure. From the training set we randomly selected groups of five compounds. Each group was left out and that group was predicted by the model developed from the remaining observations. This process was carried out 100 times.

It is important that the model is quite stable to the inclusion–exclusion of compounds as measured by values of LOO and L5O correlation coefficients. The results of predictions on the LOO ($R_{CV,LOO}^2 = 0.65$) and L5O ($R_{CV,L5O}^2 = 0.70$) cross-validation test illustrated the quality of the obtained model.

The model was further validated by applying the Y-randomization. Several random shuffles of the Y vector were performed and the results are shown in Table 6. The low R^2 and R_{CV}^2 values show that the good results in our original model are not due to a chance correlation or structural dependence of the training set.

The extrapolation method was applied to the compounds that constitute the test set. The leverages for all the 17 compounds were computed (Table 7). None of the 17 compounds fell outside the domain of the model (warning leverage limit 0.48).

4 Conclusion

Our results lead to the conclusion that the inhibition of caspase-3 enzymes can be successfully modeled with physicochemical constants and structural descriptors. The validation procedures utilized in this work (separation of the data into two independent sets, Y-randomization) illustrates the accuracy and robustness of the produced models not only by calculating their fitness on sets of training data, but also by testing the predicting abilities of the models. The proposed method, due to the high predictive ability, could be a useful aid to the costly and time-consuming experiments for determining inhibition of caspase-3 [22, 32]. Furthermore, the produced models could be used to screen existing databases or virtual libraries in order to identify novel potent compounds. In this case, the applicability domain will serve as a valuable tool to filter out “dis-similar” compounds.

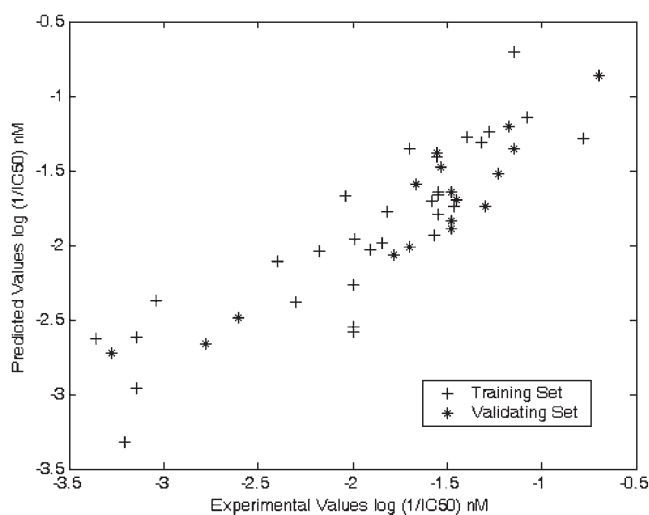


Figure 1. Experimental vs. predicted $\log(1/IC_{50})$ values for the training and validation sets.

Table 6. R^2 and R_{CV}^2 values after several Y-randomization test.

Iteration	R^2	R_{CV}^2
1	0.13	0.00
2	0.08	0.00
3	0.25	0.07
4	0.42	0.26
5	0.32	0.11
6	0.49	0.32
7	0.19	0.05
8	0.14	0.03
9	0.21	0.06
10	0.06	0.00

Table 7. Leverages for the test set.

Compound Id	Leverages (limit 0.48)
3	0.14
7	0.12
9	0.18
10	0.09
12	0.08
15	0.04
18	0.21
20	0.08
22	0.17
23	0.08
25	0.16
28	0.06
31	0.16
33	0.27
41	0.14
42	0.09
43	0.12

Acknowledgements

A. A. wishes to thank Cyprus Research Promotion Foundation (Grant No. PENEK/ENISX/0603/05) and A. G. Leventis Foundation for its financial support. G. M. thanks the Empirikion Foundation for financial support.

References

- [1] A. Afantitis, G. Melagraki, K. Makridima, A. Alexandridis, H. Sarimveis, O. Igglessi-Markopoulou, *J. Mol. Struct. – THEOCHEM* **2005**, *716*, 193–198.
- [2] M. Karelson, V. S. Lobanov, A. R. Katritzky, *Chem. Rev.* **1996**, *96*, 1027–1044.
- [3] G. Melagraki, A. Afantitis, H. Sarimveis, O. Igglessi-Markopoulou, C. T. Supuran, *Bioorg. Med. Chem.* **2006**, *14*, 1108–1114.
- [4] D. K. Agrafiotis, V. S. Lobanov, F. R. Salemme, *Nat. Rev. Drug Discov.* **2002**, *1*, 337–346.
- [5] K. L. Bhat, S. Hayik, L. Sztandera, C. W. Bock, *QSAR Comb. Sci.* **2005**, *24*, 831–843.
- [6] T. Netzeva, A. O. Aptula, S. H. Chaudary, J. C. Duffy, T. W. Schultz, G. Schüürmann, M. T. D. Cronin, *QSAR Comb. Sci.* **2003**, *22*, 575–582.
- [7] G. S. Salvesen, V. M. Dixit, *Cell* **1997**, *91*, 443–446.
- [8] P. Villa, S. H. Kaufmann, W. C. Earnshaw, *Trends Biochem. Sci.* **1997**, *22*, 388–393.
- [9] N. A. Thornberry, Y. Lazebnik, *Science* **1998**, *281*, 1322–1326.
- [10] S. Kothakota, T. Azuma, C. Reinhard, A. Klipel, J. Tang, K. Chu, T. McGarry, M. Kirschner, K. Kothe, D. Kwiatkowski, L. Williams, *Science* **1997**, *278*, 294–298.
- [11] N. Hoglen, L. Chen, C. Fisher, B. Hirakawa, T. Groessl, P. Contreras, *J. Pharmacol. Exp. Ther.* **2004**, *309*, 634–640.
- [12] B. Han, D. Xu, J. Choi, Y. Han, S. Xanthoudakis, S. Roy, J. Tam, J. Vaillancourt, J. Colucci, R. Siman, A. Giroux, G. Robertson, R. Zamboni, D. Nicholson, D. Holtzman, *J. Biol. Chem.* **2002**, *277*, 30128–30136.
- [13] A. Haunstetter, S. Izumo, *Circ. Res.* **1998**, *82*, 1111–1129.
- [14] C. Wellington, M. Hayden, *Clin. Genet.* **2000**, *57*, 1–10.
- [15] C. Hansch, B. Bonavida, A. Jazirehi, J. Cohen, C. Milliron, A. Kurup, *Bioorg. Med. Chem.* **2003**, *11*, 617–620.
- [16] C. Hansch, A. Jazirehi, S. Mekapati, R. Garg, B. Bonavida, *Bioorg. Med. Chem.* **2003**, *11*, 3015–3019.
- [17] S. Cai, L. Guan, S. Jia, Y. Wang, W. Yang, B. Tseng, J. Drewe, *Bioorg. Med. Chem. Lett.* **2004**, *14*, 5295–5300.
- [18] Y. Wang, L. Guan, S. Jia, W. Yang, B. Tseng, J. Drewe, S. X. Cai, *Bioorg. Med. Chem. Lett.* **2005**, *15*, 1379–1383.
- [19] Y. Wang, J.-C. Huang, Z.-L. Zhou, W. Yang, J. Guastella, J. Drewe, S. X. Cai, *Bioorg. Med. Chem. Lett.* **2004**, *14*, 1269–1272.
- [20] R. Todeschini, V. Consonni, R. Mannhold, H. Kubinyi, H. Timmerman (Eds.), *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, **2000**.
- [21] S. Wold, L. Eriksson, in: Van de Waterbeemd, H., (Ed.), *Statistical Validation of QSAR Results, Chemometrics Methods in Molecular Design*, VCH Weinheim, **1995**, 309–318.
- [22] A. Tropsha, P. Gramatica, V. K. Gombar, *QSAR Comb. Sci.* **2003**, *22*, 1–9.
- [23] A. Golbraikh, A. Tropsha, *J. Mol. Graph. Model.* **2002**, *20*, 269–276.

- [24] M. Shen, C. Beguin, A. Golbraikh, J. Stables, H. Kohn, A. Tropsha, *J. Med. Chem.* **2004**, *47*, 2356–2364.
- [25] A. C. Atkinson, *Plots, Transformations and Regression*, Clarendon Press, Oxford (UK), **1985**, p. 282.
- [26] R. Mannhold, A. Petrauskas, *QSAR Comb. Sci.* **2003**, *22*, 466–475.
- [27] W. P. A. Walters, M. A. Murcko, *Curr. Opin. Chem. Biol.* **1999**, *3*, 384–387.
- [28] J. Devillers, (Ed.), *Comparative QSAR*, Taylor and Francis, Washington, DC, **1998**.
- [29] C. Hansch, A. Leo, in: S. Heller (Ed.), *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*, ACS, Washington, DC, **1995**.
- [30] R. W. Kennard, L. A. Stone, *Technometrics* **1969**, *11*, 137–148.
- [31] W. Wu, B. Walczak, D. L. Massart, S. Heuerding, F. Erni, I. R. Last, K. A. Prebble, *Chemometr. Intell. Lab. Syst.* **1996**, *33*, 35–46.
- [32] A. O. Aptula, N. G. Jeliaskova, T. W. Schultz, M. T. D. Cronin *QSAR Comb. Sci.* **2005**, *24*, 385–396.