

QSAR Models for Prediction of PPAR δ Agonistic Activity of Indanylacetic Acid Derivatives

Viney Lather, Miguel X. Fernandes*

Centro de Química da Madeira, Departamento de Química da Universidade da Madeira, Campus da Penteada, 9000-390 Funchal, Portugal, E-mail: mxf@uma.pt

Keywords: Artificial neural network (ANN), Indanylacetic acids, Multiple linear regression (MLR), Peroxisome proliferator activated receptor (PPAR), QSAR

Received: July 23, 2008; Accepted: January 11, 2009

DOI: 10.1002/qsar.200810092

Abstract

Peroxisome Proliferator Activated Receptor β/δ (PPAR β/δ), one of three PPAR isoforms is a member of nuclear receptor superfamily and ubiquitously expressed in several metabolically active tissues such as liver, muscle, and fat. Tissue specific expression and knock-out studies suggest a role of PPAR δ in obesity and metabolic syndrome. Specific and selective PPAR δ ligands may play an important role in the treatment of metabolic disorders. Indanylacetic acid derivatives reported as potent and specific ligands against PPAR δ have been studied for the Quantitative Structure–Activity Relationships (QSAR). Molecules were represented by chemical descriptors that encode constitutional, topological, geometrical, and electronic structure features. Four different approaches, *i.e.*, random selection, hierarchical clustering, k-means clustering, and sphere exclusion method were used to classify the dataset into training and test subsets. Forward stepwise Multiple Linear Regression (MLR) approach was used to linearly select the subset of descriptors and establish the linear relationship with PPAR δ agonistic activity of the molecules. The models were validated internally by Leave One Out (LOO) and externally for the prediction of test sets. The best subset of descriptors was then fed to the Artificial Neural Networks (ANN) to develop non-linear models. Statistically significant MLR models; with R^2 varying from 0.80 to 0.87 were generated based on the different training and test set selection methods. Training of ANNs with different architectures for the different training and test selection methods resulted in models with R^2 values varying from 0.83 to 0.94, which indicates the high predictive ability of the models.

1 Introduction

Dysregulation of Fatty Acid (FA) levels is the characteristic of some of the most prevalent medical disorders, including obesity, cardiovascular disease, and type 2 diabetes. Drugs like thiazolidinediones and fibrates reduce elevated levels of circulating FAs mediating their effects by binding to the Peroxisome Proliferator Activated Receptors (PPARs) [1, 2]. Three closely related mammalian PPAR subtypes (α , γ , and δ) have been identified. PPARs regulate the expression of target genes related to the carbohydrate and lipid metabolism. Although the biology of PPAR δ is the least well understood of the PPARs, this subtype activates transcription through the same type of response elements as PPAR α and PPAR γ and presumably also modulates lipid and/or glucose homeostasis [3, 4].

The first proposed pharmacological role for PPAR δ was in the regulation of cholesterol homeostasis [5, 6]. Apart from the effects on cholesterol homeostasis and glycemic control, PPAR δ activation has been suggested to attenuate inflammation and slow the progression of atherosclerosis by direct and indirect mechanisms [7]. Human polymorphism studies further suggested that PPAR δ is involved in cholesterol metabolism [8]. Based on the above findings regarding the role, PPAR δ plays in regulating different biological mechanisms, there is a considerable interest in creating novel PPAR δ agonists from both scientific and clinical points of view. X-ray crystallography and structure-based drug design approaches are being used by different research groups to design new PPAR δ agonists [9, 10]. Recently, Markt *et al.* [11] used a ligand-based pharmacophore modeling approach for the parallel screening

of PPAR ligands against different PPAR receptor subtypes based on the developed pharmacophore models.

Quantitative Structure–Activity Relationship (QSAR) provides drug designers with valuable information which can be used to improve the efficacy of drugs [12]. The 2D and 3D structural descriptors describing the molecular structure in terms of its constitution, geometry, reactivity, connectivity, spatial arrangement, *etc.* are calculated, and validated models can be used to forecast the approximate activity of the designed molecules.

The insulin sensitizing activity for 51 indanylacetic acid analogs, as agonists of PPAR δ , was evaluated by Wickens *et al.* [13], but the authors did not establish a QSAR model for this set of compounds. Since, a QSAR model is an imperative to improve the efficacy of compounds, Multiple Linear Regression (MLR) with forward stepwise feature selection approach has been used to generate the linear models on these 51 indanylacetic acid analogs to predict their PPAR δ agonistic activity. The same subset of descriptors was then studied to recognize the nonlinear relationships with the dependent variables by feeding to the Multi-layer Perceptrons (MLPs) neural networks. As one of the most important steps in a QSAR study is the generation of training and test sets, four different approaches: random selection approach, tree-based approach hierarchical clustering, k-means clustering, and sphere exclusion method have been used to observe the effect of these classification methods on the predictability of the resulting models. The developed models are expected to be valuable in the rational design of chemical modifications of PPAR δ ligands in order to identify potential candidates as lead structures.

2 Methodology

2.1 Dataset for Analysis

The 51 analogs of indanylacetic acid that were selected for the model development were assigned with experimental EC₅₀ values (Fret δ assay) as the indicators of their *in vitro* PPAR δ receptor agonistic activity [13]. The basic structures for these analogs are shown in Figure 1 and various substituents are enlisted in Table 1 along with their agonistic activity for PPAR δ (pEC₅₀) which was used as the dependent variable in the model development. Dataset was subdivided into training and test sets based on four different selection methods to test the effect of these selection methods on the predictive abilities of the resulting models.

2.2 Computational Details

The structures of compounds were pre-optimized using the molecular mechanics force field; MMFF included in Omega version 2.2.1, Open Eye Scientific Software, Inc [14]. Being the (S) isomer of A₁ more active as reported by Wickens *et al.* (S) enantiomers were selected for the compounds A₁–A₄ and B₆–B₈ for further studies. The optimized geometries of the parent molecule A₁ from Omega were compared with cocrystallized PPAR δ ligand (PDB ID: 1Y0S) using a shape and chemical similarity matching algorithm; Rapid Overlay of Chemical Structures (ROCS), version 2.3.1, Open Eye Scientific Software, Inc. [14, 15]. ROCS is a shape-based superposition method. Molecules are aligned by a solid-body optimization pro-

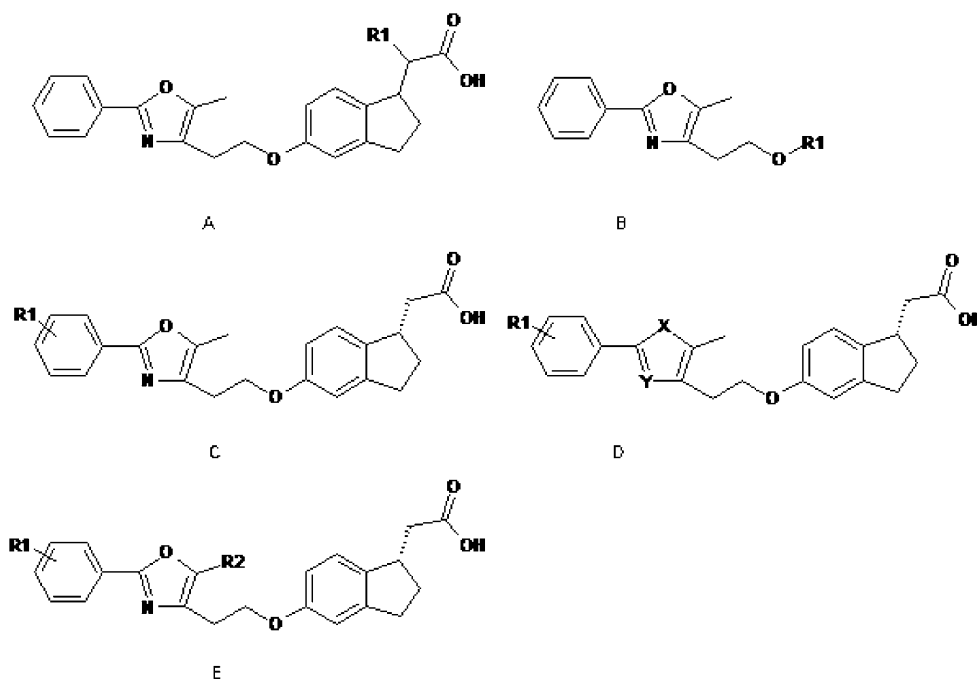


Figure 1. Basic structure of indanylacetic acid analogs.

Table 1. Dataset used for MLR and ANN QSAR analysis with corresponding experimental and predicted activities for random selection method.

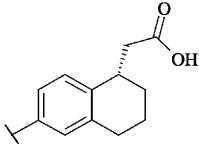
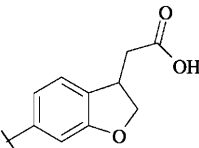
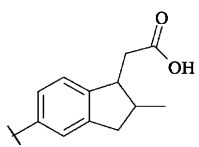
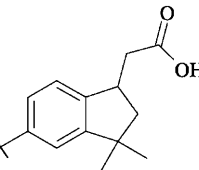
Compound no.	R1	X	Y	R2	Experimental pEC ₅₀	Predicted pEC ₅₀		Residuals	
						MLR	FFNN	MLR	FFNN
A ₁	H	—	—	—	7.57	7.60	7.48	0.03	-0.09
^a A ₂	Me	—	—	—	6.92	5.92	5.83	-1.00	-1.09
A ₃	Et	—	—	—	6.23	6.45	6.10	0.22	-0.13
A ₄	OMe	—	—	—	5.36	5.57	5.45	0.21	0.09
B ₅		—	—	—	5.50	5.92	5.58	0.42	0.08
B ₆		—	—	—	5.00	5.46	5.06	0.46	0.06
B ₇		—	—	—	6.05	5.81	6.10	-0.24	0.05
^a B ₈		—	—	—	5.19	—	6.84	2.16	1.65
C ₉	H	—	—	—	7.85	7.78	7.67	-0.07	-0.18
C ₁₀	4-MeO	—	—	—	8.52	8.50	8.54	-0.02	0.02
C ₁₁	3-MeO	—	—	—	7.74	7.99	7.96	0.25	0.22
C ₁₂	4-Et	—	—	—	8.52	8.97	8.67	0.45	0.15
^a C ₁₃	4- <i>t</i> -Bu	—	—	—	7.68	8.21	8.43	0.53	0.75
C ₁₄	4- <i>i</i> -Pr	—	—	—	9.00	8.25	8.38	-0.75	-0.62
C ₁₅	3-F	—	—	—	7.89	7.80	7.82	-0.09	-0.07
C ₁₆	4-F	—	—	—	8.15	7.74	8.06	-0.41	-0.09
C ₁₇	4-Ph	—	—	—	8.22	8.01	7.96	-0.21	-0.26
C ₁₈	4-Me	—	—	—	8.40	8.25	8.36	-0.15	-0.04
C ₁₉	3-Me	—	—	—	7.19	7.66	7.18	0.47	-0.01
C ₂₀	4-CN	—	—	—	7.46	7.62	7.49	0.16	0.03
C ₂₁	3-CN	—	—	—	7.14	7.25	6.94	0.11	-0.20
^a C ₂₂	3-Cl	—	—	—	7.89	8.04	7.50	0.15	-0.39
C ₂₃	4-Cl	—	—	—	8.52	8.27	8.47	-0.25	-0.05
^a D ₂₄	H	S	N	—	7.96	7.80	8.01	-0.16	0.05
^a D ₂₅	4-Me	S	N	—	9.10	7.77	8.25	-1.33	-0.85
D ₂₆	3,4-OCH ₂ O	S	N	—	8.30	7.75	8.22	-0.55	-0.08
D ₂₇	4-MeO	S	N	—	8.40	7.72	8.21	-0.68	-0.19
^a D ₂₈	3-MeO	S	N	—	7.21	7.27	7.36	0.06	0.15
D ₂₉	4- <i>i</i> Pr	S	N	—	8.30	8.22	8.14	-0.08	-0.16
D ₃₀	4-F	S	N	—	8.15	7.75	8.08	-0.40	-0.07
D ₃₁	3-F	S	N	—	7.33	7.54	7.77	0.21	0.44
D ₃₂	2-F	S	N	—	6.94	6.47	6.94	-0.47	0.00
^a D ₃₃	4-Cl	S	N	—	8.82	8.53	8.83	-0.29	0.01
D ₃₄	4-EtO	S	N	—	7.52	7.95	8.15	0.43	0.63
D ₃₅	3-Me	S	N	—	6.57	7.18	6.86	0.61	0.29
D ₃₆	3-CF ₃	S	N	—	7.52	8.21	7.65	0.69	0.13

Table 1. (cont.)

Compound no.	R1	X	Y	R2	Experimental pEC ₅₀	Predicted pEC ₅₀		Residuals	
						MLR	FFNN	MLR	FFNN
D ₃₇	4-CF ₃ O	S	N	—	8.70	8.50	8.58	-0.20	-0.12
D ₃₈	4-Ph	S	N	—	7.89	8.51	7.91	0.62	0.02
D ₃₉	4-Ph	N	NMe	—	7.55	7.37	7.48	-0.18	-0.07
^a D ₄₀	4-Ph	NMe	N	—	6.51	5.95	6.63	-0.56	0.12
D ₄₁	4-Et	N	NMe	—	7.51	7.36	7.62	-0.15	0.11
D ₄₂	4-Et	NMe	N	—	6.46	6.49	6.52	0.03	0.06
D ₄₃	4-MeO	N	NMe	—	7.07	6.83	6.86	-0.24	-0.21
D ₄₄	4-MeO	NMe	N	—	5.00	4.65	4.91	-0.35	-0.09
D ₄₅	H	N	NMe	—	6.59	7.31	7.07	0.72	0.48
^a D ₄₆	H	NMe	N	—	5.00	4.55	5.16	-0.45	0.16
E ₄₇	H	—	—	Et	8.15	7.51	7.37	-0.64	-0.78
E ₄₈	4-Me	—	—	Et	8.30	7.76	8.31	-0.54	0.01
E ₄₉	4-Et	—	—	Et	7.96	8.25	8.32	0.29	0.36
E ₅₀	H	—	—	Pr	8.00	8.20	8.17	0.20	0.17
E ₅₁	H	—	—	PhEt	6.22	6.32	6.37	0.10	0.15

^a Test set molecules.

cess that maximizes the overlap volume between them. Volume overlap in this context is a Gaussian-based overlap parameterized to reproduce hard-sphere volumes. ROCS uses only the heavy atoms of a ligand, hydrogens are ignored. The final geometries and quantum chemical data were obtained by subjecting all the molecules to energy refinement with semiempirical method AM1 using the AMPACTM 8 program [16]. All geometries and electronic parameters were calculated in vacuum. The following sets of keywords were used in all quantum computations: AM1 PRECISE VECTORS BONDS PI KPOLAR ENPART.

The Omega structure files and the AMPAC output files were used as an input in CODESSA program [17, 18] for the calculation of a total of 562 structural descriptors. CODESSA computes five classes of structural descriptors: constitutional, topological, geometrical, electrostatic, and quantum-chemical. The 3D optimized structures from Omega were also used as an input in E-Dragon [19, 20], an online descriptor calculating program for the calculation of topological, charge, WHIM, BCUT, and GETAWAY descriptors. In total, more than 1000 molecular descriptors were generated that were too many to be fitted in the QSAR models. Selection of descriptors was performed to reduce the pool of descriptors by eliminating those that satisfied at least one of the following conditions [21]: (i) the descriptor has a constant value for all the molecules investigated, (ii) the descriptors with a correlation coefficient less than 0.3 with the dependent variable (pEC₅₀) were regarded as a redundant, (iii) in the monoparametric correlation with (pEC₅₀), the descriptor has a squared correlation coefficient lower than 0.1, (iv) in the monoparametric correlation the descriptor has a *t*-test value lower than 0.1, (v) in the monoparametric correlation the descriptor has an *F*-test value lower than 1 at a probability

level of 0.05, (vi) highly correlated descriptors provide approximately identical information, if their pair wise correlation coefficient exceeded 0.75. After these steps, the number of descriptors was reduced to 27.

2.3 Training and Test Set Selection

Four different approaches were used for the selection of training and test sets of compounds: random selection, hierarchical clustering, K-means clustering, and sphere exclusion method. Based on the random selection method, dataset was classified into training and test subsets of 41 and 10 compounds, respectively keeping in view, the training/test set ratio of 80:20.

Topological and constitutional descriptors calculated using CODESSA program were used as the basis for calculating the similarity between the molecules for the hierarchical and K-means clustering approaches. Euclidean distance was used as the similarity measure parameter for these clustering methods, carried out using R-program version 2.6.1 [22]. Based on the hierarchical clustering approach, all the compounds in the dataset were grouped into ten clusters. A total of ten molecules were selected for the test set by randomly picking one molecule from each of the clusters. A similar procedure was followed for selecting the test set using K-means clustering.

The fourth method used for the training and test set classification was sphere exclusion method implemented in QSAR-Plus software, V-Life Sciences. The program employs the following algorithm: (i) select a point and include it in the training set; (ii) build a sphere with radius *R* with a center in this point; (iii) include all points within the sphere, except for the center, in the test set; (iv) discard all points in the sphere from the initial set; (v) if no points are

left, stop, otherwise go to step (i). The most active compound in the dataset is selected as the starting point for building a sphere. Constitutional and topological descriptors from CODESSA were used as the similarity measures. A probability value of 0.5, which defines the dissimilarity among the descriptors, was used for this method.

A further validation study of the QSAR models was carried out by generating ten different test sets for random, K-means clustering, and hierarchical clustering methods. Models were generated using the training sets and validated by the test set of ten molecules each. For K-means clustering and hierarchical clustering, a different data point has been included from each cluster in each of the test set. The resulting statistical parameters are reported as an average in terms of R^2 (ten-fold validation) and R^2_{test} (ten-fold validation).

All the four selection methods were carried out to satisfy the criteria described by Valkova *et al.* [23] and Golbraikh and Tropsha [24]: (i) diversity of the training set, which is necessary condition for building a QSAR equation applicable to further compounds of interest in the same chemical domain; (ii) closeness of the representative points of both the training and test set in the descriptor space that ensures a proper validation of the model.

2.4 QSAR-I: Forward Selection and Multiple Linear Regression Modeling

Forward feature selection with MLR was used to establish the first type of QSAR models. Using F value for the analysis of variance, R^2 and RMSE of training set as a criteria of selection, subsets of descriptors were examined for establishing the best linear QSAR. The size of descriptor subset used for model establishment was increased until no improvement was seen as well as keeping in view that the number of compounds in the training set should not be smaller than five times the number of descriptors. Variance-covariance matrices were calculated for each of the descriptors in all of the resulting linear models and the descriptors which had multicollinearity, were discarded. Tolerance and Variation Inflation Factor (VIF) were chosen as the parameters for determining the collinearity among the variables. VIF values of 1–4 indicate the non-collinearity among the variables. Among the remaining models after the elimination process, the one that had the minimum RMSE was chosen as the best. The goodness of the regression fits were estimated using parameters, such as R^2 , RMSE, q^2 [Leave One Out (LOO) cross-validation], and F -statistics. After model development with randomly selected set of training compounds, the best model was further examined by the test set molecules. The same subset of descriptors used in the best model (random selection) was then used for the generation of models for the other training set selection methods. The Z -score method was adopted for the detection of outliers in the training as well as test sets. Z -score can be defined as the absolute differ-

ence between the value of the model and the activity field, divided by the square root of the mean square error of the dataset. Any compound which has a value of Z -score close to 2.5 during the generation of a particular QSAR model was considered as an outlier.

2.5 QSAR-II: Forward Selection and Artificial Neural Network Modeling

ANN calculations were performed with Statistica version 7.0, Statsoft Inc. The best subset of descriptors selected in QSAR-I was fed to neural networks to develop QSAR-II models. The neural networks used in this study were three-layer fully connected MLPs (feed forward neural networks). Such networks are supposed to identify the nonlinear relationship between the structural descriptors and biological activity of the molecules. The networks were trained using the training set molecules.

Each neuron in the network was connected to all neurons in neighboring layer(s) through adjustable weights. Network training is the process of adjusting the weights, such that the error is minimized, and the number of input layer neurons is equal to the number of descriptors. The descriptors values as well as the pEC_{50} values were linearly scaled in order to ensure that some descriptors are not weighed more heavily than others due to their nature. Network properties such as the number of hidden layer neurons, learning rate, and number of iterations were optimized using both RMS errors of training and test sets as the selection criteria. Z -score method was used for the detection of outliers, if any, in the training and test sets for all the methods.

3 Results and Discussion

3.1 Conformational Analysis

Crystallographic data from PDB database of the bound PPAR δ subtype ligands shows that the ligands adopted a Y-shaped conformation in the ligand binding domain of the PPAR δ receptor. To confirm that the geometries/lowest energy conformations of indanylacetic acid derivatives obtained from Omega adopted Y-shape conformations, shape of the indanylacetic acids were superposed with those of the co-crystallized 1Y0S ligand using ROCS. Figure 2 shows the superposition of shape of the parent compound A_1 in the dataset with that of 1Y0S ligand. Close correspondence of the Omega optimized conformation of these molecules with co-crystallized data validated the selection of these geometries for the calculation of 3D descriptors by CODESSA and E-Dragon.

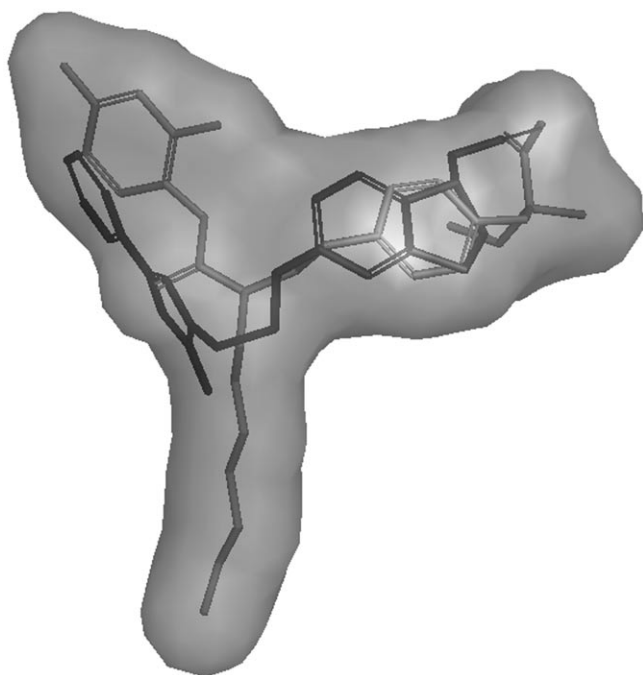


Figure 2. ROCS overlay of parent molecule A_1 against cocrystallized ligand 1Y0S. Molecule A_1 is shown in black; the molecular surface is that of 1Y0S ligand.

3.2 QSAR-I

Linear regression modeling with forward feature selection methodology was carried out. Figure 3 shows the change in the statistics of R^2 and Standard Error of Estimate (SEE), as the number of descriptors is increased from 1 to 7 in the regression models. It can be observed that as the number of descriptors is increased from 1 to 7, regression coefficient, R^2 , values increased. A further increase in the number of descriptors in the regression models did not result in any further increase in R^2 . A similar trend can be seen for the SEE value which decreases as the number of descriptors is increased from 1 to 7. The best MLR model for the randomly selected training set of compounds contained seven descriptors which are related to the dependent variable as follows:

$$\text{pEC}_{50} = -67.438 + 58.685\text{QC1} + 49.322\text{QC2} + 4.248\text{QC3} + 2.025\text{BEHm6} - 9.120\text{GM1} - 0.020\text{T(N...O)} - 0.714\text{QC4}$$

$$n = 41; R^2 = 0.856; q^2 = 0.78; F = 28.03; \text{RMSE} = 0.387$$

Table 2 shows the details about these descriptors. Pair wise correlation for these descriptors with pEC_{50} ranged from 0.3 to 0.62. Table 3 shows the inter-correlation matrix of the seven descriptors with the biological activity, and Table 4 shows the statistical details of the resulting model. All the t values are significant with low p -values which confirm the significance of each descriptor. The F statistic

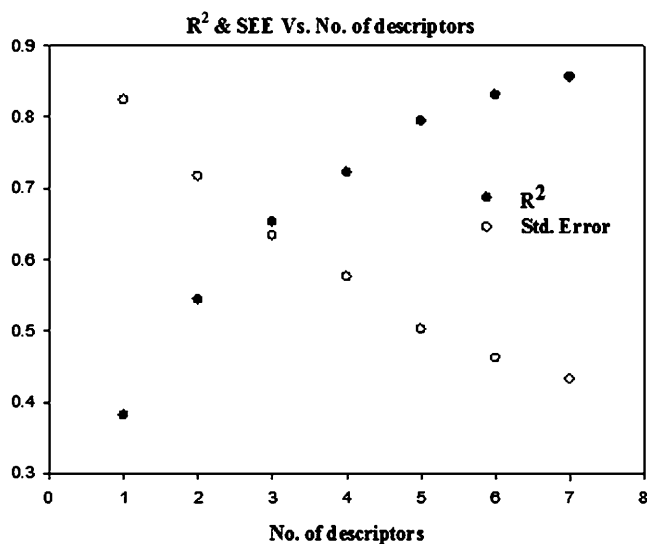


Figure 3. Plot of R^2 and SEE versus the number of descriptors for an MLR model (random selection method).

for this model is 28.03 (compared to the critical value of 2.42 at the 0.05 level of significance) with a p -value of less than 0.000, and the lowest partial F value for the coefficient was 20.5. Furthermore the tolerance and VIF values are all on the lower side, which indicates the absence of multicollinearities in the model. Thus the model is considered to be statistically valid. The RMSE for the training set was 0.387. The predictive ability of the model was validated internally by LOO method resulting in a q^2 value of 0.78, which indicates that the model is highly predictive. The model was further validated externally by the test dataset of ten compounds with a resulting R^2_{pred} value of 0.54 and RMSE of 0.91. However, after the removal of compound 8 (residual = 2.35, Z-score = 2.4) detected as an outlier, from the test dataset, the R^2_{pred} value increased significantly to 0.81 and RMSE decreased to 0.64. Figure 4 shows the fit plot of experimental versus predicted pEC_{50} values for the training as well as the test sets. The resulting model was further validated by dividing the dataset into 10 different training and test sets of 41 and 10 compounds each. An average R^2 of 0.86 and 0.71 was determined for the training and test set, respectively after removal of outlier, wherever found, based on Z-score method. Since, the descriptors with greater coefficients are more determining in regression equations, it can be concluded that according to the statistics of this model, the most important descriptor is QC1 (ESP-minimum net atomic charge for an O atom), followed by the QC2 (Minimum atomic orbital electronic population), whereas the least determining descriptor is T(N...O). ESP-minimum net atomic charge for an O atom is an electrostatic potential-based charge calculated descriptor. This descriptor reflects the charge distribution for the oxygen atoms of the molecule and characterizes the intermolecular electrostatic interactions. Positive coefficient

in QSAR equation indicates that an increase in the activity is observed with an increase in the minimum net atomic charge on O atom. The charges derived from the electrostatic potential have the advantage that they are physically more meaningful than Mulliken's charges [27] and this procedure for the charge calculation is of particular relevance in simulations of intermolecular interaction, especially describing molecules with biological activity [28]. The charges at the oxygen atoms reflect H-bond interactions, other electrostatic interactions and obviously play an important role for the biological effect of the studied compounds. Minimum atomic orbital electronic population (Min-OP) for a given atomic species in the molecule is a simplified index to describe the electrophilic ability of the molecule and connected to the hydrogen donor capabilities of the molecule. BEHm6, a BCUT index is highest eigenvalue no. 6 of the Burden matrix weighted by atomic mass. BCUT descriptors are the eigenvalues of a modified connectivity matrix known as the Burden matrix. The low *t*-value for BEHm6 tells about the lower significance of this descriptor in the model, but on inclusion in the model through forward selection methods, it leads to a high change in the R^2 statistics as can be seen in Figure 2. QC3 (minimum total interaction for a C–O bond) is the minimum total interaction energy for a C–O bond and is calculated as the summation of two terms, the minimum electronic exchange energy for a C–O bond and the minimum Coulomb interaction energy for a C–O bond; it may be related to the conformational changes or atomic reactivity in the molecule. These descriptors may be related to the formation of highly reactive radical centers in the aromatic systems. T(N...O) is the summation of topological distances between N and O atoms in the molecule. A negative coefficient value indicates that increasing the distance between the two atoms in a molecule increases the biological activity. GM1 (XY shadow/XY rectangle) is a geometrical descriptor characterizing the shape and extent of the molecule in terms of its 3D coordinates. This descriptor represents a two-dimensional projection on the X–Y plane of a three-dimensional molecule. Orientation of a molecule along the axes of inertia (X-coordinate) casts a shadow of the molecule projected on the X–Y plane. Normalized shadows are calculated by XY shadow/XY rectangle. A negative coefficient shows that the activity increases with decrease in the value of XY shadow, which means that a

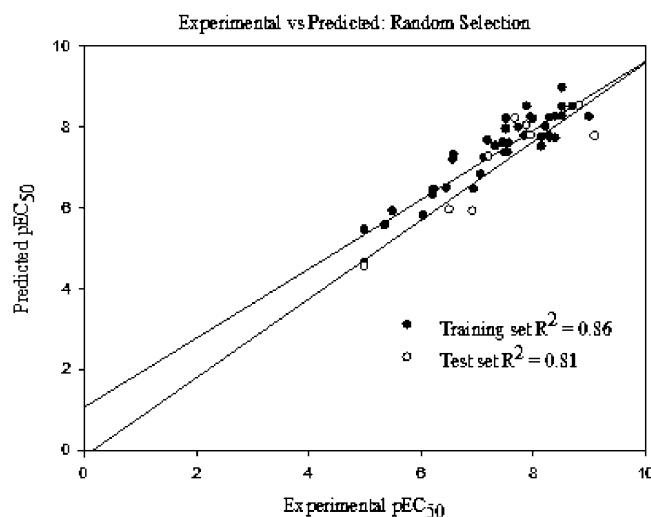


Figure 4. Plots of predicted activity by MLR against the experimental activity for the data used in MLR selection method.

smaller area of molecular shadow in the enclosing rectangle will benefit the activity.

The lowest Unoccupied Molecular Orbital (LUMO) energy is obtained from molecular orbital calculations and represents the electron accepting capability of the molecule and characterizes the susceptibility of the molecule toward attack by the nucleophile or governs the way the molecule interacts with receptor. Molecules with low lying LUMO are more able to accept the electron than those with higher energy. Trend in indanylacetic acid derivatives was such that the lower the LUMO energy of the molecule, higher the activity of the molecule, which suggests the role of H-bond in the ligand–receptor interactions.

The same subset of descriptors was then used to build the QSAR models for the other training and test sets selection methods in order to see the change in the predictability of the models, when different training and test compounds are chosen. By means of K-means clustering, the dataset was divided into training set of 41 compounds, for which the MLR model based on the descriptors used in random selection method, resulted in an R^2 of 0.863 with an SEE of 0.43, and other statistics details are shown in Table 5. The model was validated by prediction for the external test set of ten compounds ($R^2=0.467$). However after

Table 2. Descriptors selected by models for the prediction of PPAR δ agonistic activity.

No.	Name	Description	Type	Reference
1	QC1	ESP- Min net atomic charge for an O atom	Quantum-chemical	[25]
2	QC2	Min. atomic orbital electronic population	Quantum-chemical	[25]
3	BEHm6	Burden eigenvalue weighted by atomic mass	BCUT index	[26]
4	QC3	Min. total interaction for a C-O bond	Quantum-chemical	[25, 29]
5	T (N...O)	Sum of topological distances between N and O atoms	Topological	[30]
6	GM1	XY shadow/XY rectangle	Geometrical	[31]
7	QC4	E_{LUMO}	Quantum-chemical	[25]

Table 3. Correlation matrix for the inter-correlation of structural descriptors and their correlation with the activity.

	pEC ₅₀	QC1	QC2	QC3	QC4	BEHm6	GM1	T (N...O)
pEC ₅₀	1.000							
QC1	0.618	1.000						
QC2	0.442	-0.065	1.000					
QC3	0.397	0.282	-0.037	1.000				
QC4	-0.357	-0.082	-0.085	-0.037	1.000			
BEHm6	0.350	0.273	-0.281	-0.077	-0.354	1.000		
GM1	-0.442	-0.271	-0.118	-0.211	-0.153	-0.343	1.000	
T(N...O)	-0.300	0.038	-0.092	0.108	-0.277	-0.066	-0.255	1.000

Table 4. Statistics for the best MLR model for random selection method.

Descriptor	Coefficient	SE	<i>t</i>	Significance (<i>p</i>)	Tolerance	VIF
Constant	-67.438	21.625	-3.119	0.001	-	-
QC1	58.685	10.680	5.490	0.000	0.823	1.215
QC2	49.322	9.760	5.054	0.000	0.754	1.327
BEHm6	2.025	1.300	1.553	0.010	0.541	1.849
QC3	4.248	1.180	3.576	0.001	0.879	1.138
T(N...O)	-0.020	0.005	-4.355	0.000	0.850	1.177
GM1	-9.120	2.570	-3.548	0.001	0.627	1.594
QC4	-0.714	0.295	-2.418	0.010	0.693	1.443

the removal of compound 8 (residual = 2.15, Z-score = 2.4) detected as an outlier, from the test dataset, the R^2_{pred} value increased significantly to 0.68 (compared to standard value of 0.6 for the test sets) and RMSE value of 0.64. Further validation based on the ten test sets generated for the molecules in the dataset resulted in R^2 of 0.86 and 0.74, respectively for the training and test set.

The dataset was divided into a training set of 41 compounds using hierarchical clustering, for which the MLR model based on the same subset of descriptors resulted in an R^2 value of 0.78 with an SEE value of 0.56. The outlier (compound 8) was included in the training set for the hierarchical clustering. Upon exclusion of the outliers based on the Z-score method, the resulting model ($R^2=0.86$) was used for the prediction of test set compounds. Compounds 8 and 25 were detected as the outliers. The R^2_{pred} value for an external test set of ten compounds was found to be 0.62, which validated this QSAR model. A ten-fold validation of the QSAR model resulted in average R^2 of 0.86 and 0.63 for the training and test sets, respectively. Statistical results are shown in Table 5.

Table 5. MLR models statistics of training/test set selection methods.

Method	R^2	RMSE (training)	q^2	<i>F</i>	<i>p</i> -value	R^2 (test)	R^2 (ten-fold validation)	R^2_{test} (ten-fold validation)
Random selection	0.86	0.39	0.78	28.03	0.000	0.81	0.86	0.71
K-means clustering	0.86	0.38	0.80	29.68	0.000	0.68	0.86	0.74
Hierarchical clustering	0.86	0.37	0.67	27.97	0.000	0.62	0.86	0.63
Sphere exclusion method	0.80	0.40	0.70	16.48	0.000	0.83	-	-

For the sphere exclusion method, dataset were classified into a training set of 40 and a test set of 11 compounds. The MLR model for the training set of compounds resulted in an R^2 of 0.70 with an SEE of 0.59. Upon removal of outliers (compound 8 and 25) based on Z-score, the resulting model ($R^2=0.80$) was then used for the external test set. The R^2_{pred} value for the external test set of compounds was found to be 0.83.

3.3 QSAR-II

The seven descriptors which were selected by the QSAR model I were then used to establish the nonlinear models. Networks used were of three-layered type, containing a bias neuron in each layer and a single neuron in the output layer. MLP networks were trained with the training set of compounds using a back propagation algorithm followed by conjugate gradient descent in the second phase. The values of each input was normalized between $[-1, 1]$, to bring the values of input variables into the dynamic range of the sigmoid transfer function in the ANN. The weights were initialized to a uniformly-distributed random value,

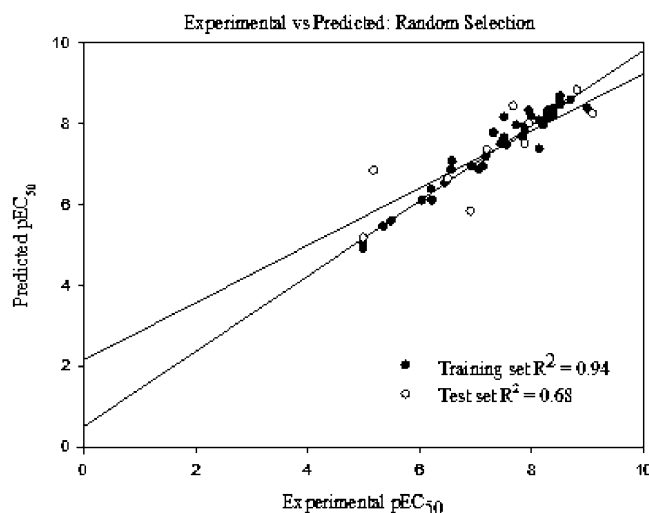
Table 6. Statistics of the ANN models for training and test sets.

Method	Network architecture	R^2 (training)	R^2 (test)	RMSE (training)	RMSE (test)	R^2 (ten-fold validation)	R^2_{test} (ten-fold validation)
Random selection	7-4-1	0.94	0.68	0.25	0.73	0.93	0.67
K means clustering	7-5-1	0.94	0.61	0.25	0.79	0.93	0.65
Hierarchical clustering	7-6-1	0.93	0.88	0.27	0.57	0.93	0.80
Sphere exclusion method	7-6-1	0.83	0.83	0.37	0.50	–	–

within a range whose minimum and maximum values are given by the minimum/mean and maximum/SD fields, respectively, and ranged between -1 and 1 . Learning rate was set at 0.01 in the start and momentum parameter set at 0.3 . To avoid the overfitting and overtraining in the predictability by neural networks, training set was subdivided into a learning set and into a selection set. The first set was used to train the network, whereas the selection set was used to monitor the training process. The optimal training endpoint and network architecture were determined on the basis of this selection set. The network architecture and training endpoint which gave the lowest RMSE, for the prediction of the selection set was then used for further study. To ensure that the results obtained were not due to chance, the predictions were repeated several times with different initial weights. The optimal training ANN endpoint required 100 training epochs for backpropagation and 500 for conjugate gradient descent algorithms. The results of the models for the different training and test set selection methods are shown in Table 6.

The MLP networks were trained with same dataset of 41 compounds for the random classification method used in QSAR-I models. The optimal network architecture was found to be a 7-4-1 model. The plot of the experimental *versus* predicted pEC_{50} for the training and test sets of random classification method are shown in Figure 5. The R^2 values for the training and test sets were 0.94 and 0.68 , respectively. The RMSE values for the training and test sets were 0.25 and 0.73 , respectively. Compared with the linear models, an improved predictive performance was observed through the ANN approaches. The activity for compound 8 was predicted poorly by the nonlinear model, probably because it could not be detected as an outlier based on Z-score, but still the overall prediction of the test set was high when compared to the linear model. A ten-fold validation of the ANN model resulted in an average R^2 of 0.93 and 0.67 for the training and test sets, respectively.

A similar approach for training of the networks was used for the other training and test sets selection methods. The neural network model for the k-means clustering method resulted in R^2 values of 0.94 and 0.61 for the training and test sets, respectively. For the hierarchical clustering method, the model resulted in R^2 values of 0.93 for the training set and 0.88 for the test set (upon removal of the outlier). The model for the sphere exclusion classification method resulted in R^2 values of 0.83 (upon removal of out-

**Figure 5.** Plots of predicted activity by ANN against the experimental activity for the data used in random selection method.

lier 8 based upon Z-score) for both the training and test sets.

Analysis of the linear and ANN models shows that nonlinear prediction of PPAR δ agonistic activity, using the same subset of descriptors, was better than their counterpart from linear models. This shows that there is a complex relationship between the structure and activity of compounds. It is not possible simply by inspection to determine the influence of one input variable on the output variable in the case of ANN models. ANN models have the advantages of solving complex relationship between the structure and activity, but a loss of transparency occurs for these models.

The results of the linear regression models presented in this work clearly indicate that the electrostatic interactions have a decisive role in determining the PPAR δ agonistic activity of the indanylacetic acid derivatives. Compounds substituted α to the acetic acid, *e.g.*, A_2 – A_4 have lower value of minimum net atomic charge on O atom and lower value of minimum atomic orbital electronic population and are less active. A similar effect was observed for the compounds B_5 – B_8 . The low activity for these molecules can also be due to the lower values observed for the descriptor; minimum total interaction energy for the C–O bond, which can be a measure of the conformational

changes and atomic reactivity of these molecules. Amongst the molecules C_9-C_{23} , $D_{24}-D_{46}$, and $D_{47}-D_{51}$, the agonistic activity is more influenced by descriptors like BEHm6, XY shadow and LUMO energy. Molecules $C_{24}-C_{46}$; derivatives of 2-phenylthiazoles and 2-phenyl-1-methyl-imidazoles are inactive when the values of XY shadow is higher or BEHm6 is lower. The trend shows that size and shape plays a major role for this series as explained by XY shadow or BEHm6. Larger molecules with 4-phenyl substitutions are inactive. Similar effect was seen for compounds with 1-methyl imidazoles. Higher XY shadow values for these compounds made them inactive.

Further evaluation of the QSAR models and physical significance of the descriptors involved lead to identify the features like unsubstituted indanylacetic acid is essential moiety for the biological activity. Replacing or substituting the indanyl moiety, resulting in a change in the electrostatic properties as explained by the quantum chemical descriptors, leads to a decrease or loss of the activity. Distance between the indanylacetic acid moiety and oxazole ring as explained by the T(N...O) and BEHm6 descriptors showed that increasing the chain length results in poor activity. Furthermore, replacing the oxazole/thiazole moiety by 1-methyl-imidazole evident by high XY shadow/XY rectangle values resulted in loss of the potency. Replacing 5-methyl group on oxazole by ethyl or propyl groups produced no changes in the quantum or other descriptors and maintained the activity.

4 Conclusions

The MLR and ANN regression analyses were employed to study the PPAR δ agonistic activity of indanylacetic acid derivatives in order to develop interpretable and predictive QSAR models. Different methods were used to select the training and test sets in order to study the effects of these methods on the predictability of the resulting models. The models based on random selection, K-means clustering, and hierarchical clustering had similar predictive abilities for the training as well as test sets. The QSAR models generated using the sphere exclusion selection method were found to be highly predictive for the test sets for both the MLR and ANN models. The results showed that predictability of the models can be influenced by the training and test set selection methods, although all the resulting models were found to be statistically valid. ANN models were found to be slightly more successful than MLR in predicting agonistic activity, reflecting a nonlinear relationship between the molecular descriptors and the PPAR δ agonistic activity for this set of molecules. However, ANN models do not allow a clear interpretation of descriptor contributions, as is available from the linear model. Overall, these models (particularly the ANN with hierarchical clustering and sphere exclusion selection method) can be successfully used to speed up the design and devel-

opment of novel PPAR δ ligands, and provide a much needed treatment for several of the most prevalent disorders like obesity and metabolic syndrome.

Acknowledgements

We thank Fundação para a Ciência e Tecnologia (Portugal) for grant SFRH/BPD/30954/2006 attributed to Viney Lather.

References

- [1] A. L. Catapano, *Pharmacol. Res.* **1992**, *26*, 331–340.
- [2] B. M. Spiegelman, *Diabetes* **1998**, *47*, 507–514.
- [3] A. Schimdt, N. Endo, S. J. Rutledge, R. Vogel, D. Shinar, G. A. Rodan, *Mol. Endocrinol.* **1992**, *6*, 1634–1641.
- [4] S. A. Kliewer, B. M. Forman, B. Blumber, E. S. Ong, U. Borgmeyer, D. J. Mangelsdorf, K. Umersono, R. M. Evans, *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 7355–7359.
- [5] M. D. Leibowitz, C. Fievet, N. Hennuyer, J. Peinado-Onsurbe, H. Duez, J. Berger, C. A. Cullinan, C. P. Sparrow, J. Baffic, G. D. Berger, C. Santini, R. W. Marquis, R. L. Tolman, R. G. Smith, D. E. Moller, J. Auwerx, *FEBS Lett.* **2000**, *473*, 333–336.
- [6] P. Sauerberg, G. S. Olsen, L. Jeppsen, J. P. Mogensen, I. Pettersson, C. B. Jeppesen, J. R. Daugaard, E. D. Galsgaard, L. Ynddal, J. Fleckner, V. Panajotova, Z. Polivka, P. Pihera, M. Havranek, E. M. Wulff, *J. Med. Chem.* **2007**, *50*, 1495–1503.
- [7] C. H. Lee, A. Chawla, N. Urbiztondo, D. Liao, W. A. Boisvert, R. M. Evans, *Science* **2003**, *302*, 453–457.
- [8] J. Skogberg, K. Kannisto, T. N. Cassel, A. Hamsten, P. Eriksson, E. Ehrenborg, *Arterioscler. Thromb. Vasc. Biol.* **2003**, *23*, 637–643.
- [9] J. Kasuga, I. Nakagome, A. Aoyama, K. Sako, M. Ishizawa, M. Ogura, M. Makishima, S. Hirono, Y. Hashimoto, H. Miyachi, *Bioorg. Med. Chem.* **2007**, *15*, 5177–5190.
- [10] R. Epple, M. Azimioara, R. Russo, B. Bursulaya, S. S. Tian, A. Gerken, M. Iskandar, *Bioorg. Med. Chem. Lett.* **2006**, *16*, 2969–2973.
- [11] P. Markt, D. Schuster, J. Kirchmair, C. Laggner, T. Langer, *J. Comput. Aided Mol. Des.* **2007**, *21*, 575–590.
- [12] Y. C. Martin, *Quantitative Drug Design*, Marcel Dekker, New York **1978**.
- [13] P. Wickens, C. Zhang, X. Ma, Q. Zhao, J. Amatruda, W. Bullock, M. Burns, L. D. Cantin, C. Y. Chuang, T. Claus, M. Dai, F. D. Cruz, D. Dickson, F. J. Ehr Gott, D. Fan, S. Heald, M. Hentemann, C. I. Iwuagwu, J. S. Johnson, E. Kumarsinghe, D. Ladner, R. Lavoie, S. Liang, J. N. Livingston, D. Lowe, S. Magnuson, G. Mannelly, I. Mugge, H. Ogutu, S. Pleasic-Williams, R. W. Schoenleber, J. Shapiro, T. Shelekhin, L. Sweet, C. Town, M. Tsutsumi, *Bioorg. Med. Chem. Lett.* **2007**, *17*, 4369–4373.
- [14] OEChem, Version 1.3.4, OpenEye Scientific Software, Inc., Santa Fe, NM, USA, www.eyesopen.com.
- [15] T. S. Rush, III, J. A. Grant, L. Mosyak, A. Nicholls, *J. Med. Chem.* **2005**, *48*, 1489–1495.
- [16] AMPAC 8, © 1992–2004, Semichem, Inc. PO Box 1649, Shawnee, KS66222.
- [17] A. R. Katritzky, A. Oliferenko, A. Lomaka, M. Karelson, *Bioorg. Med. Chem. Lett.* **2002**, *12*, 3453–3457.

- [18] CODESSA 2.7, © 1992–2004, Semichem, Inc. PO Box 1649, Shawnee, KS66222.
- [19] I. V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. A. Palyulin, E. V. Rodchenko, N. S. Zefirov, A. S. Makarenko, V. Y. Tanchuk, V. V. Prokopenko, *J. Comput. Aided Mol. Des.* **2005**, *19*, 453–463.
- [20] VCCLAB, Virtual Computational Chemistry Laboratory, <http://www.vcclab.org>, 2005.
- [21] T. Ivanciuc, O. Ivanciuc, *Internet Electron. J. Mol. Des.* **2002**, *1*, 94–107.
- [22] R Development Core Team, 2004. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.
- [23] I. Valkova, M. Vracko, S. C. Basak, *Anal. Chim. Acta* **2004**, *509*, 179–186.
- [24] A. Golbraikh, A. Tropsha, *Mol. Div.* **2000**, *5*, 231–234.
- [25] M. Karelson, V. S. Lobanov, *Chem. Rev.* **1996**, *96*, 1027–1043.
- [26] F. R. Burden, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225–227.
- [27] U. C. Singh, P. A. Kollman, *J. Comput. Chem.* **1984**, *5*, 129–145.
- [28] D. E. Williams, *J. Comput. Chem.* **1994**, *15*, 719–732.
- [29] Y. Ren, H. Liu, S. Li, X. Yao, M. Liu, *Bioorg. Med. Chem. Lett.* **2007**, *17*, 2474–2482.
- [30] J. Devillers, A. T. Balaban, *Topological Indices and Related Descriptors in QSAR and Drug Design*, Gordon & Breach, New York (NY) **2002**.
- [31] J. E. Code, K. E. Perko, D. M. Yourtee, A. J. Holder, E. Kostoryz, *J. Biomater. Sci., Polym. Ed.* **2007**, *18*, 1457–1474.