Monatsh Chem (2010) 141:111–118
DOI 10.1007/s00706-009-0225-4

ORIGINAL PAPER

Quantitative structure–activity relationship study of nonpeptide antagonists of CXCR2 using stepwise multiple linear regression analysis

**Jahan B. Ghasemi · Parvin Zohrabi ·
Habibollah Khajehsharifi**

Received: 21 September 2009 / Accepted: 20 November 2009 / Published online: 19 December 2009

Abstract The chemokine receptor CXCR2 plays an important role in recruiting granulocytes to sites of inflammation and has been proposed as an important therapeutic target. A linear quantitative structure–activity relationship model is presented for modeling and predicting biological activities of CXCR2 antagonists. The model was produced by using the multiple linear regression technique on a database that consists of 55 nonpeptide antagonists of CXCR2. Stepwise regression as a variable selection method was used to develop a regression equation based on 43 training compounds, and predictive ability was tested on 12 compounds reserved for that purpose. Appropriate models with low standard errors and high correlation coefficients were obtained. The mean effect of descriptors and standardized coefficients shows that the mean atomic van der Waals volume is the most important property affecting the biological activities of the molecules. The square regression coefficient of prediction set for the multiple linear regression method was 0.912.

Keywords QSAR · Chemokine receptor · CXCR2 antagonists · Multiple linear regression · Biological activity · Molecular modeling

J. B. Ghasemi (✉)
Chemistry Department, Faculty of Sciences,
K.N. Toosi University of Technology, Tehran, Iran
e-mail: Jahan.Ghasemi@gmail.com

P. Zohrabi
Chemistry Department, Faculty of Sciences,
Razi University, Kermanshah, Iran

H. Khajehsharifi
Chemistry Department, Faculty of Sciences,
Yasouj University, Yasouj, Iran

Introduction

Chemokines are a set of small proteins, typically comprising 70–80 amino acids, which play an important role in the recruitment and activation of inflammatory cells. They may be classified according to the nature of conserved cysteine motifs, and generally fall into two main categories, the CC and CXC chemokines. Both subfamilies include a number of potent chemoattractants and activators of different leukocyte subsets [1]. One family of chemokines is characterized by the presence of an intervening amino acid between the first pair of conserved cysteines and is known as the _Cys-X-Cys_ (CXC) or a-chemokine family [2]. CXC chemokines that contain the sequence Glu-Leu-Arg (ELR) before the first N-terminal cysteine residue mediate, in part, the recruitment of neutrophils and a subset of monocytes. ELR + chemokines act through CXC chemokine receptors CXCR1 and CXCR2 [3, 4].

CXCR1 binds IL-8 and granulocyte chemotactic protein-2 (GCP-2/CXCL6) with high affinity, whereas CXCR2 is promiscuous, binding seven known ELR a-chemokines with high affinity, including GRO-a (CXCL-1), GRO-b (CXCL-2), GRO-c (CXCL-3), epithelial neutrophil activating peptide 78 (ENA-78/CXCL-5), GCP-2 (CXCL-6), neutrophil-activating peptide-2 (NAP-2/CXCL-7), and IL-8 (CXCL-8). Furthermore, CXCR2 is expressed by a wide range of cell types, for example neutrophils, mast cells, T cells, keratinocytes, and cerebellar neurons [5–7].

CXCR2 mouse gene knockout studies show that there are elevated leukocytes and lymphocytes without apparent pathogenic consequences, indicating that CXCR2 is not required for normal physiology [8]. Increased levels of CXCR2 and its ligand IL-8 have been observed in humans with diseases such as arthritis, asthma, rheumatoid arthritis, psoriasis, reperfusion injury, and chronic obstructive pulmonary disease (COPD) [9]. This suggests that the CXCR2 receptor and IL-8 may play a pivotal role in these inflammatory disorders. Therefore, antagonists of CXCR2 receptor could, in principle, be used for treatment of inflammatory and related diseases. CXCR2 antagonists have indeed attracted much attention as targets for small-molecule drug discovery in the last decade [10].

One of the most successful approaches to the prediction of chemical properties starting with molecular structural information only is modeling of quantitative structure–activity/property relationships (QSAR/QSPR). The QSAR models provide significant additional insight into the relationship between molecular structure and fundamental processes and phenomena in chemistry [11]. Quantitative structure–activity relationships (QSAR) are mathematical equations relating chemical structure to a wide variety of physical, chemical, biological, and technological properties [12]. A major step in constructing QSAR models is finding a set of molecular descriptors that represent variation in the structural properties of the molecules. A wide variety of descriptors have been reported for use in QSAR analysis [13, 14].

The advantage of this approach over other methods is that it requires a knowledge of chemical structure only and is not dependent on any experimental properties. Construction of QSAR models is essential for understanding the molecular mechanism of action of receptor antagonists, their design, and virtual screening [15]. Currently, several QSAR models utilizing a flexible docking approach have been shown to be highly efficient in the description of ligand–receptor interactions [16].

The main objective of this work was to develop an accurate, simple, fast, and less expensive method for calculation of pIC_{50} for a set of nonpeptide antagonists of CXCR2 using theoretical molecular descriptors. In this work a QSAR study was performed to develop models that relate the structures of 55 CXCR2 antagonists to their biological activities. The stepwise multiple linear regression (MLR) using SPSS (ver 11.5) as variable selection software was used to model biological activity with the structural descriptors.

Results and discussion

External validation

For regression analysis the data set was separated into two groups: training and prediction sets. The molecules included in these sets were selected randomly. The training set, consisting of 43 molecules, was used for model generation using the SPSS software package. The prediction set, consisting of 12 molecules, was used to evaluate the generated model.

The predictive power of the regression model developed on the basis of the selected training set is estimated from the values predicted for prediction set chemicals, by use of the external Q^2 that is defined as [17]:

$$Q^2_{\text{ext}} = 1 - \frac{\sum_{i-1}^{\text{pred}} (y_i - \hat{y}_i)^2}{\sum_{i-1}^{\text{pred}} (y_i - \bar{y}_{\text{tr}})^2} \tag{1}$$

where y_i and $\hat{y}_i$ are the measured and predicted (over the prediction set) values of the dependent variable, respectively, and $\bar{y}_{\text{tr}}$ is the averaged value of the dependent variable for the training set; the summations cover all the compounds in the prediction set [18]. Other measures used to define the accuracy of prediction of the proposed QSARs are the square of the correlation coefficient (R^2) calculated for the prediction chemicals by applying the model developed on the training set, and the root-mean-squared error of prediction (RMSEP). The RMSEP is a measurement of the average difference between predicted and experimental values in the prediction step. The orthogonality of the descriptors in the model was established through variance inflation factor (VIF) [19, 20]. The VIF is defined as $1/(1 - r_i^2)$ where r_i is the multiple correlation coefficient for the ith variable regressed on the $p - 1$ others, p being the number of variables contributing to the model. A VIF value larger than 5 indicates that the information of the descriptors may be hidden by the correlation of the other descriptors [21].

Regression models

Multiple linear regression analysis has been carried out to derive the best QSAR model. The MLR technique was performed on the molecules of the training set. After regression analysis, a few suitable models were obtained among which the best model was selected; this is presented in Eq. 2. The developed model was then used to predict the pIC_{50} values of the compounds in the test set, which have not been used for the model development. MLR analysis provided a useful equation that can be used to predict the pIC_{50} of drugs based upon these parameters. This QSAR model for the biological activities of the CXCR2 antagonists includes four molecular descriptors.

$$\begin{aligned} \text{p}IC_{50} = {} & -9.7250 + 4.5742 \times \text{GATS5}v + 0.0536 \\ & \times \text{RDF065}m + 3.3776 \times R3u + 9.2569 \times V\text{m} \end{aligned} \tag{2}$$

The statistical characteristics of the best four indices in the MLR model are shown in Table 1. The orthogonality of the descriptors (VIF) in the MLR model is in agreement with the limit. The standardized regression coefficient reveals the significance of an individual descriptor presented in the regression model. The greater the

Table 1 Model parameter values and standardized coefficients for the MLR model

Source	Model parameter		Mean effect	VIF
	Value	SE		
Intercept	−9.72505	3.177059		
GATS5v	4.574166	0.613444	3.628058	1.146417
RDF065m	0.053591	0.017797	0.520429	1.175182
R3u	3.377594	0.919507	5.532106	1.440488
V_m	9.256878	2.866415	6.527176	1.422166

absolute value of a coefficient, the greater the weight of the variable in the model. Figure 1 shows that the effect of the mean atomic van der Waals volume (V_m), R autocorrelation of lag 3/unweighted (R3u), Geary autocorrelation-lag 5 weighted by atomic van der Waals volume (GATS5v), and the radial distribution function at 6.5 Å interatomic distances weighted by atomic mass and contributed negatively (RDF065m) are more significant than the other descriptors. It is worthy to note that the V_m descriptor has the highest significance for the MLR model.

Evaluation of the regression model

Model validation techniques are needed in order to distinguish between true and random correlations and to estimate the predictive power of the model. The real predictive ability of any QSAR model cannot be judged solely by using internal validation, it has to be validated on the basis of predictions for pIC_{50} of CXCR2 antagonists not included in the training set. For evaluation of the predictive power of the constructed model, the optimized model was used for prediction of the pIC_{50} values of 12 CXCR2 antagonists in the prediction set, which were not used in the optimization procedure.

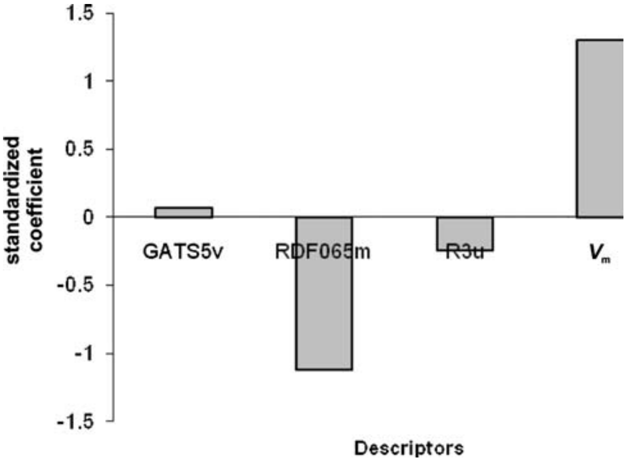


Fig. 1 Standardized coefficients versus descriptors in the MLR model

Table 2 Experimental pIC_{50}, predicted pIC_{50}, residuals values, and relative error for external prediction set by the MLR method

No.	pIC_{50} (Exp.)	pIC_{50} (Pred.)	Residuals	RE (%)
29	5.114	5.127	0.013	0.254
31	5.456	5.658	0.202	3.709
34	5.638	5.619	−0.019	−0.334
38	5.854	5.431	−0.423	−7.229
13	6.066	5.664	−0.402	−6.624
12	6.495	5.861	−0.634	−9.762
53	6.77	6.386	−0.384	−5.666
4	6.943	7.136	0.193	2.782
2	7.201	7.501	0.300	4.171
23	7.495	7.180	−0.315	−4.201
9	7.658	7.761	0.103	1.344
16	8.032	8.232	0.200	2.491

In Table 2, the predicted values of pIC_{50} obtained by the MLR method and the percent relative errors of prediction are presented. Plots of predicted pIC_{50} versus experimental pIC_{50} and the residuals (predicted pIC_{50} – experimental pIC_{50}) versus experimental pIC_{50} values, obtained by the MLR modeling, are shown in Figs. 2 and 3, respectively. The agreement observed between the predicted and experimental values in Fig. 2 and the random distribution of residuals about zero mean in Fig. 3 confirms the good predictive ability of MLR modeling. For the constructed model, general statistical parameters were selected to evaluate the prediction ability of the model for pIC_{50}. The statistical parameters root mean squares error of prediction (RMSEP, measures the precision in prediction), relative error of prediction (REP), standard error of residual in

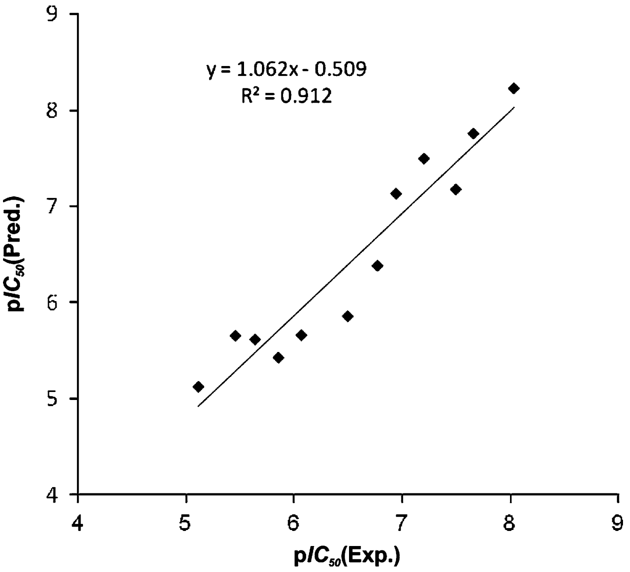


Fig. 2 pIC_{50} values predicted by MLR modeling versus experimental pIC_{50} values

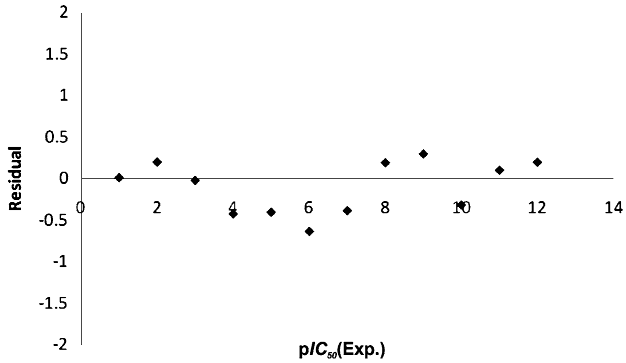


Fig. 3 Residual versus experimental pIC_{50} in the MLR model

prediction (SEP, measures the precision in prediction), and squared regression coefficient calculated for the MLR model are listed in Table 3. The data presented in this table indicate that validation has good statistical qualities with low prediction errors. Table 4 presents the correlation matrix, where it is clear that the four selected descriptors are highly decorrelated.

Interpretation of the descriptors

The best four-parameter equation for prediction of pIC_{50} for an unknown compound included V_m (constitutional), R3u (GETAWAY descriptor), GATS5v (2D-autocorrelation descriptor), and RDF065m (RDF descriptor).

Comparison of the mean effects of the descriptors appearing in the MLR model shows that the V_m (mean atomic van der Waals volume) of the molecules has the largest effect on the pIC_{50} of the CXCR2 antagonists (Table 1). The mean effect of a descriptor is the product of its mean and the regression coefficient in the MLR model [22]. V_m is a constitutional descriptor. Constitutional descriptors are basically related to the number of atoms and bonds in each molecule. These are the most simple and commonly used descriptors, the most common constitutional descriptors are molecular weight, van der Waals volume, atomic electronegativities and polarizabilities, number of atoms, non-H atoms, covalent bonds, multiple bonds, bond orders, aromatic ratio, number of double and triple bonds, aromatic bonds, and different types of (n-membered) rings and benzene-like rings. These descriptors are insensitive to any conformational change, do not distinguish among isomers, and are either 0D descriptors or 1D descriptors [18]. V_m is a bulk property, which describes the size of a molecule.

Table 3 Statistical parameters obtained by applying the MLR method to the test set

Parameter	RMSEP	SEP	REP (%)	R^2_{pred}	Q^2_{ext}
Value	0.317343	0.331454	4.910057	0.912	0.8770217

Table 4 Correlation matrix for MLR model

	pIC_{50}	GATS5v	RDF065m	R3u	V_m
pIC_{50}	1				
GATS5v	0.721101	1			
RDF065m	0.547172	0.324529	1		
R3u	0.191318	−0.05648	0.105711	1	
V_m	0.128573	−0.02947	0.068998	−0.5196	1

The second descriptor in the models is the R3u; R3u is a GETAWAY descriptor (3D). The GETAWAY (geometry, topology, and atom-weights assembly) descriptors [13, 23] are recently proposed molecular descriptors derived from a new representation of molecular structure, the molecular influence matrix (MIM), denoted by **H** and defined as the following

$$\mathbf{H} = \mathbf{M}\left(\mathbf{M}^{\mathrm{T}} \times \mathbf{M}\right)^{-1}\mathbf{M}^{\mathrm{T}} \tag{3}$$

where **M** is the molecular matrix constituted by the centered Cartesian coordinates x, y, and z of the molecule atoms (hydrogens included) in a chosen conformation, and the superscript T refers to the transposed matrix.

The diagonal elements h_{ii} of the MIM, called leverages, encode atomic information and represent the "influence" of each molecule atom in determining the whole shape of the molecule. In fact, mantle atoms always have higher h_{ii} values than atoms near the molecule center. Moreover, the magnitude of the maximum leverage in a molecule depends on the size and shape of the molecule itself. Each off-diagonal element h_{ij} represents the degree of accessibility of the jth atom to interactions with the ith atom or, in other words, the attitude of the two considered atoms to interact with each other. A negative sign for the off-diagonal elements means that the two atoms occupy opposite molecular regions with respect to the center, hence the degree of their mutual accessibility should be low.

Two sets of theoretically closely related molecular descriptors have been devised: H-GETAWAY descriptors have been calculated from the MIM **H**, and R-GETAWAY descriptors are from the influence/distance matrix **R** where the elements of the MIM are combined with those of the geometry matrix.

The influence/distance matrix **R** is the new symmetric A × A molecular matrix, proposed here, whose elements resemble the single terms in the sums of the gravitational indices, defined as the following:

$$[\boldsymbol{R}]_{ij} \equiv \left[\frac{\sqrt{h_{ii} \times h_{jj}}}{r_{ij}}\right]_{ij} \quad i \neq j \tag{4}$$

where h_{ii} and h_{jj} are the leverages of the two considered atoms and r_{ij} is their geometric distance. The row sums of the influence/distance matrix encode some useful information that could be related to the presence of significant substituents or fragments in the molecule. In fact, it has been observed that larger row sums correspond to terminal atoms that are located very close to other terminal atoms such as those in substituents on a parent structure. With the objective of catching relevant chemical information, these new descriptors have been defined by applying: traditional matrix operators, concepts of the theoretical information and spatial autocorrelation formulas, and weighting the molecule atoms accounting atomic mass, polarizability, van der Waals volume, and electronegativity. R3u is R autocorrelation of lag 3/unweighted.

Another descriptor in the MLR model is GATS5v; GATS5v belongs to 2D-autocorrelation descriptors (2D). This set consists of 96 descriptors calculated from the molecular graph by summing the products of atom weights of the terminal atoms of all the paths of the considered path length (the lag). The 2D-autocorrelations by Moreau–Broto (ATS), Moran (MATS), and Geary (GATS) algorithms are calculated from lag 1 to lag 8 for four different weighting schemes [24–26],

$$\mathrm{ATS}(p_k, l) = \sum_i \delta_{ij} p_{ki} p_{kj} \tag{5}$$

$$\mathrm{MATS}(p_k, l) = \frac{N}{2L} \frac{\sum_{ij} \delta_{ij}(p_{ki} - \bar{p}_k)(p_{kj} - \bar{p}_k)}{\sum_i (p_{ki} - \bar{p}_k)} \tag{6}$$

$$\mathrm{GATS}(p_k, l) = \frac{(N-1)}{4L} \frac{\sum_{ij} \delta_{ij}(p_{ki} - \bar{p}_k)(p_{kj} - \bar{p}_k)}{\sum_i (p_{ki} - \bar{p}_k)} \tag{7}$$

where ATS (p_k,l), MATS (p_k,l), and GATS (p_k,l) are Moreau–Broto's autocorrelation coefficient, Moran's index, and Geary's coefficient at spatial lag l, respectively, p_{ki} and p_{kj} are the values of property k of atoms, i and j, respectively, $\bar{p}_k$ is the average value of property k, L is the number of nonzero values in the sum, N is the number of atoms in the molecule, and δ (l,d_{ij}) is a Dirac-delta function defined as:

$$\delta(l, d_{ij}) = \left\{ \begin{matrix} 1 & \text{if} & d_{ij=1} \\ 0 & \text{if} & d_{ij\neq 1} \end{matrix} \right\} \tag{8}$$

where d_{ij} is the topological distance or spatial lag between atoms i and j.

Spatial autocorrelation measures the level of interdependence between properties, and the nature and strength of that interdependence. In a molecule, Moran's and Geary's spatial autocorrelation analysis tests whether the value of an atomic property at one atom in the molecular structure is independent of the values of the property at neighboring atoms. If dependence exists, the property is said to exhibit spatial autocorrelation. The autocorrelation vectors represent the degree of similarity between molecules. Four different weighting schemes have been used: atomic masses (m), atomic van der Waals volumes (v), atomic Sanderson electronegativities (e), and atomic polarizabilities (p). Autocorrelation vectors were calculated for spatial lags l ranging from 1 up to 8. The autocorrelation descriptors are denoted by the scheme: type of descriptor-spatial lag-weighting property; GATS5v is the Geary autocorrelation-lag 5 weighted by atomic van der Waals volume. The definition of autocorrelation-like indices is a very active field of research covering from small drugs to proteins, which has generated some recent publications including research articles and reviews [27, 28].

The last descriptor in the MLR model which has the smallest mean effect was the RDF065m, these RDF descriptors belonging to the class of radial distribution function descriptors are based on the distance distribution in the geometrical representation of the molecule. In addition, the RDF also provides valuable information about bond distances, ring types, planar and non-planar systems, atom types, and other important structural motifs. The RDF code has been proved to be a good representation of the 3D structure which has several merits, for example independence of the number of atoms; unambiguity regarding the three-dimensional arrangement of the atoms, and invariance against translation and rotation of the entire molecule.

The RDF of an ensemble of N atoms can be interpreted as the probability distribution of finding an atom in a spherical volume of radius r. The RDF used in this work is as follows:

$$g(r) = f \sum_{1}^{N-1} \sum_{i \succ j}^{N} A_i A_j e^{-B\left(r - r_{ij}\right)^2} \tag{9}$$

$$f = 1 \Big/ \sqrt{\sum_r [g(r)]^2} \tag{10}$$

where f is a scaling factor, N is the number of atoms, A is the atomic properties of atoms i and j, B is smoothing parameter that defines the probability distribution of the individual distances, r_{ij} is the distance between the atoms i and j, and g (r) was calculated at a number of discrete points with defined intervals.

Each molecule was represented by a vector of length 32. The parameter B was set to 25 Å^{-2} corresponding to a total resolution of 0.2 Å in the defined distance r. The RDF for the structure derivations was calculated with the atomic properties. RDF065m is the radial distribution function at 6.5 Å interatomic distances weighted by atomic mass and contributed negatively [29].

Conclusion

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 reduce substantially the time and effort required to discover new medicines or improve current ones in terms of their efficacy. QSAR technology employs statistical methods to derive quantitative mathematical relationships linking chemical structure and biological activity. In this study we used MLR to model and predict pIC_{50} of 55 CXCR2 antagonists including *N*,*N*′-diphenylureas, nicotinamide *N*-oxides, quinoxalines, and triazolethiols. The MLR analysis provided a useful equation and the most relevant set of descriptors was selected by the stepwise variable selection method. The results obtained indicate that four descriptors, V_{m}, R3u, GATS5v, and RDF065m, play an important role in the biological activities of drug structures. The high correlation coefficients (0.912) and low prediction errors obtained confirm good predictive ability of the model.

Materials and methods

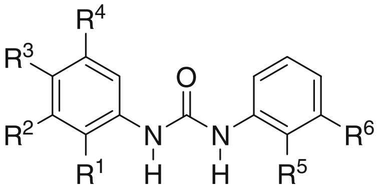
The QSAR model for estimation of the pIC_{50} of CXCR2 antagonists is established in the following steps:

- molecular structure input and generation of the files containing the chemical structures is stored in a computer-readable format;
- quantum mechanics geometry is optimized with a semi-empirical method;
- structural descriptors are computed;
- structural descriptors are selected; and
- the structure—pIC_{50} model is generated by MLR and statistical analysis.

Data set

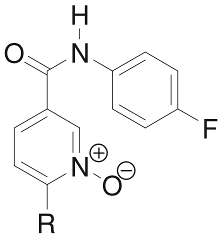
In this investigation, 55 CXCR2 antagonists were taken from the literature [30]. Four main classes of substances represented in the dataset are *N*,*N*′-diphenylureas, nicotinamide *N*-oxides, quinoxalines, and triazolethiols. The structures of the compounds investigated and their biological activities are shown in Tables 5, 6, 7, 8. The negative logarithm of the IC_{50} value [pIC_{50} or −log (IC_{50})] was adopted as a dependent variable in the QSAR analyses, with the IC_{50} values expressed in molar (M) units. The dataset was split into a training set and a testing set. The training set of 43 compounds was used to adjust the parameters of the models, and the test set of 12 compounds was used to evaluate its prediction ability.

Table 5 Structures and biological activities of the *N,N*′-diphenylureas



Compound	R^1	R^2	R^3	R^4	R^5	R^6	IC_{50} (nM)	pIC_{50}
1	OH	H	Cl	H	Br	H	906	6.043
2	OH	Cl	Cl	H	Br	H	63	7.201
3	OH	$CONH_2$	Cl	H	Br	H	10	8.000
4	OH	CH_2NH_2	Cl	H	Br	H	114	6.943
5	OH	SO_2NH_2	Cl	H	Br	H	7	8.155
6	OH	SO_2NMe_2	Cl	H	Br	H	12	7.921
7	OH	H	CN	H	Br	H	25	7.602
8	OH	Br	CN	H	Br	H	6	8.222
9	OH	Cl	CN	H	Br	H	22	7.658
10	OH	CN	Cl	H	Br	H	57	7.244
11	OH	H	NO_2	H	Br	H	22	7.658
12	OH	H	NO_2	H	H	H	320	6.495
13	OH	NO_2	H	H	H	H	860	6.066
14	OH	H	H	NO_2	H	H	10,900	4.963
15	OH	H	CN	H	H	H	200	6.699
16	OH	SO_2NH_2	Cl	H	Cl	Cl	9.3	8.032
17	–N=N–NH–		CN	H	Br	H	39	7.495

Table 6 Structures and biological activities of the nicotinamide *N*-oxides



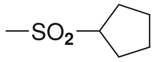
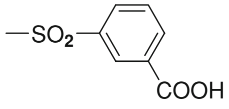
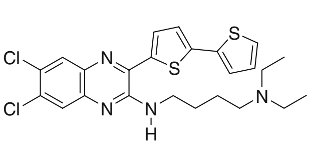
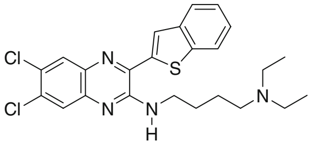
Compound	R	IC_{50} (nM)	pIC_{50}
18	$–SO_2CH_3$	130	6.886
19	$–SO_2C_2H_5$	130	6.886
20	$–SO_2CH(CH_3)_2$	400	6.398
21		460	6.337
22	$–SO_2C_6H_5$	90	7.046
23		32	7.495
24	$–SO_2CH_2C_6H_5$	280	6.553

Table 7 Structures and biological activities of the investigated quinoxalines



25



26

Compound	IC_{50} (nM)	pIC_{50}
25	160	6.796
26	30	7.553

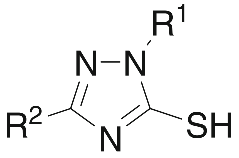
Computer hardware and software

All calculations were run on a Dell personal computer with Windows XP as operating system. ChemDraw Ultra version 9.0 (ChemOffice 2005, CambridgeSoft) software was used for drawing the molecular structures [31]. The structures of the compounds were first pre-optimized with the Molecular Mechanics Force Field (MM+) procedure implemented in HyperChem software (version 7, Hypercube) and the resulting geometries were further refined by means of the semi empirical method AM1. The optimization was preceded by the Polak–Rebiere algorithm to reach 0.42 kJ (mol Å)$^{-1}$ root mean square gradient. Molecular descriptors were calculated using the Dragon software package [32]. The software contains scripts for generating 1,497 descriptors of different types including: constitutional, topological, RDF, GETAWAY, functional groups, WHIM, Randic, 3D-Morse, etc. [18]. The software automatically eliminates constant variables in a given dataset. For descriptors with a correlation higher than 0.95, parameters are set such that only one is retained in the dataset. Final descriptor selection task was accomplished by using stepwise regression using SPSS.

Selection of descriptors

Selection of relevant descriptors, which relate the biological activities to the molecular structure, is an important step in constructing a predictive model. The calculated descriptors were collected in a data matrix **X** of dimensions ($n \times m$), where n and m are the number of molecules and descriptors, respectively. A column vector (***y***) was made from the pIC_{50} data. The stepwise regression method was used as the variable selection method to select the suitable descriptors among 164 theoretical descriptors generated by Dragon software. In stepwise regression, the first selected explanatory variable has the highest correlation with ***y***. Then, explanatory variables are consecutively added to the model in a forward selection procedure, based on their correlation with the ***y***-residuals. The significance of the model improvement is evaluated using the statistical *F* test [33] and each time a new variable is included into the model, the backward elimination step follows in which the *F* test detects variables that can be removed from the model without changing the residuals significantly. The variable selection procedure terminates, when no additional variable significantly improves the given model. By using these

Table 8 Structures and biological activities of the investigated triazolethiols



Compound	R^1	R^2	IC_{50} (nM)	pIC_{50}
27	$C_6H_5CH_2$	C_6H_5	2,400	5.620
28	3-$OHC_6H_4CH_2$	C_6H_5	4,400	5.357
29	$C_6H_5CH_2$	4-Pyridinyl	7,700	5.114
30	$C_6H_5CH_2$	2-Furanyl	4,200	5.377
31	$C_6H_5CH_2$	4-CNC_6H_4	3,500	5.456
32	$C_6H_5CH_2$	3-$CF_3C_6H_4$	3,500	5.456
33	$C_6H_5CH_2$	4-$CF_3C_6H_4$	2,800	5.553
34	$C_6H_5CH_2$	4-$CH_3OC_6H_4$	2,300	5.638
35	$C_6H_5CH_2$	3,5-diClC_6H_3	2,000	5.699
36	$C_6H_5CH_2$	2-Thienyl	2,000	5.699
37	$C_6H_5CH_2$	2-$CH_3C_6H_4$	1,400	5.854
38	$C_6H_5CH_2$	2-$CH_3OC_6H_4$	1,400	5.854
39	$C_6H_5CH_2$	3-ClC_6H_4	1,000	6.000
40	$C_6H_5CH_2$	2-FC_6H_4	890	6.051
41	$C_6H_5CH_2$	4-ClC_6H_4	830	6.081
42	$C_6H_5CH_2$	3,4-diClC_6H_3	800	6.097
43	$C_6H_5CH_2$	2,5-diClC_6H_3	670	6.174
44	$C_6H_5CH_2$	2-ClC_6H_4	450	6.347
45	$C_6H_5CH_2$	2,4-diClC_6H_3	410	6.387
46	$C_6H_5CH_2$	2-BrC_6H_4	350	6.456
47	$C_6H_5CH_2$	2,3-diClC_6H_3	350	6.456
48	4-$CH_3OC_6H_4CH_2$	2,4-diClC_6H_3	10,000	5.000
49	3-$CH_3OC_6H_4CH_2$	2,4-diClC_6H_3	4,200	5.377
50	3-$CH_3C_6H_4CH_2$	2,4-diClC_6H_3	730	6.137
51	$C_6H_5CH_2CH_2$	2,4-diClC_6H_3	450	6.347
52	4-$ClC_6H_4CH_2$	2,4-diClC_6H_3	300	6.523
53	3-$C_6H_5OC_6H_4CH_2$	2,4-diClC_6H_3	170	6.770
54	3-$ClC_6H_4CH_2$	2,4-diClC_6H_3	92	7.036
55	3-$ClC_6H_4CH_2$	2-ClC_6H_4	28	7.553

criteria, a subset of four descriptors remained, which kept the most interpretive information for pIC_{50}.

Stepwise multiple linear regression

The objective of stepwise MLR regression [33] is to construct a multivariate model for the dependent variable, $\boldsymbol{y}$, based on a few deliberately selected explanatory variables. The best equation is selected on the basis of the highest multiple correlation coefficient (r^2). The MLR method provides an equation linking the structural features to the property of the compounds for predicting the property of interest. The equation takes the following form:

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n \quad (11)$$

where y is the quantitative property or activity being predicted (dependent variable), x_i is an independent (descriptive) variable, b_0 is the intercept, and b_i is the regression coefficient for x_i. The software SPSS was used for MLR analysis. The MLR model was built using a training set and validated using an external prediction set. Multiple linear regression techniques based on least-squares procedures are very often used for estimating the coefficients involved in the model equation [34].

References

1. Walters I, Austin C, Austin R, Bonnert R, Cage P, Christie M, Ebden M, Gardiner S, Grahames C, Hill S, Hunt F, Jewell R, Lewis S, Martin I, Nicholls D, Robinson D (2008) Bioorg Med Chem Lett 18:798
2. Murphy PM, Baggiolini M, Charo IF, Hebert CA, Horuk R, Matsushima K, Miller LH, Oppenheim JJ, Power CA (2000) Pharmacol Rev 52:145
3. Hay D, Sarau H (2001) Curr Opin Pharmacol 1:242
4. Ahuja S, Murphy P (1996) J Biol Chem 271:20545
5. Baggiolini M (2001) J Intern Med 250:91
6. Lax P, Limatola C, Fucile S, Trettel F, Di BS, Renzi M, Ragozzino D, Eusebi F (2002) J Neuroimmunol 129:66
7. Liehn EA, Schober A, Weber C (2004) Arterioscler Thromb Vasc Biol 24:1891
8. Cacalano G, Lee J, Kikly K, Ryan A, Pitts-Meek S, Hultgren B, Wood I, Moore W (1994) Science 265:682
9. Bizzarri C, Allegretti M, Di Bitondo R, Neve Cervellera M, Colotta F, Bertini R (2003) Curr Med Chem 2:67
10. Busch-Petersen J (2006) Curr Top Med Chem 6:1345
11. Ghasemi J, Saaidpour S, Brown SD (2007) Theochem 805:27
12. Katritzky AR, Petrukhin R, Tatham D (2001) J Chem Inf Comput Sci 41:679
13. Consonni V, Todeschini R, Pavan M, Gramatica P (2002) J Chem Inf Comput Sci 42:693
14. Krenkel G, Castro EA, Toropov AA (2001) J Mol Struct (Theochem) 542:107
15. Godden JW, Stahura FL, Bajorath J (2004) J Med Chem 47:5608
16. Moitessier N, Henry C, Maigret B, Chapleur Y (2004) J Med Chem 47:4178
17. Carlsen L, Sørensen PB, Thomsen M (2001) Chemosphere 43:295
18. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley-VCH, Weinheim
19. Chatterjee S, Hadi A, Price B (2000) Regression analysis by examples, 3rd edn. Wiley-VCH, New York
20. Shapiro S, Guggenheim B (1998) Quant Struct Act Relat 17:327
21. Cho DH, Lee SK, Kim BT, No KT (2001) Bull Korean Chem Soc 22:388
22. Jalali-Heravi M, Konuze E (2002) Internet Electron J Mol Des 1:410
23. Consonni V, Todeschini R, Pavan M (2002) J Chem Inf Comput Sci 42:682
24. Moreau G, Broto P (1980) Nouv J Chim 4:359
25. Moran PAP (1950) Biometrika 37:17
26. Geary RF (1954) Incorp Stat 5:115
27. Agüero-Chapin G, González-Dıaz H, Molina R, Varona-Santos J, Uriarte E, González-Dıaz Y (2006) FEBS Lett 580:723
28. González-Dıaz H, Vilar S, Santana L, Uriarte E (2007) Curr Top Med Chem 7:1025
29. Hemmer MC, Steinhauer V, Gasteiger J (1999) Vib Spectrosc 19:151
30. Khlebnikov AI, Schepetkin IA, Quinn MT (2006) Bioorg Med Chem 14:352
31. ChemOffice (2005) CambridgeSoft Corporation, http://www.cambridgesoft.com/
32. Todeschini R, Milano Chemometrics, QSPR Group, http://www.disat.unimib.it/chm
33. Massart DL, Vandeginste BGM, Buydens LMC, De Jong S, Lewi PJ, Smeyers-Verbeke J (1997) Handbook of chemometrics and qualimetrics, Part A. Elsevier, Amsterdam
34. Martens H, Næs T (1989) Multivariate calibration. Wiley, Chichester