Available online at www.sciencedirect.com
ScienceDirect
European Journal of Medicinal Chemistry 44 (2009) 877–884
EUROPEAN JOURNAL OF MEDICINAL CHEMISTRY
http://www.elsevier.com/locate/ejmech

Original article

A novel QSAR model for predicting the inhibition of CXCR3 receptor by 4-*N*-aryl-[1,4] diazepane ureas

Antreas Afantitis [a,b,*], Georgia Melagraki [b,c], Haralambos Sarimveis [c], Olga Igglessi-Markopoulou [c], George Kollias [a,*]

[a] *Biomedical Sciences Research Center ''Alexander Fleming'', 34 Fleming Street, Vari 16672, Greece*
[b] *Department of ChemoInformatics, NovaMechanics Ltd, Nicosia, Cyprus*
[c] *School of Chemical Engineering, National Technical University of Athens, Athens, Greece*

Received 6 December 2007; received in revised form 12 March 2008; accepted 23 May 2008
Available online 10 July 2008

Abstract

A linear quantitative structure–activity relationship (QSAR) model is presented for modeling and predicting the inhibition of CXCR3 receptor. The model was produced by using the multiple linear regression (MLR) technique on a database that consists of 32 recently discovered 4-*N*-aryl-[1,4] diazepane ureas. The key conclusion of this study is that ^3k, ChiInf8, ChiInf0, AtomCompTotal and ClogP affect significantly the inhibition of CXCR3 receptor by diazepane ureas. The selected physicochemical descriptors serve as a first guideline for the design of novel and potent antagonists of CXCR3.

Keywords: CXCR3; Inflammatory diseases; Molecular modeling; QSAR

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 decrease substantially the time and effort required to discover new medicines or improve current ones in terms of their efficacy [1,2]. QSAR technology employs statistical methods to derive quantitative mathematical relationships linking chemical structure and biological activity [3–8].

Chemokines play a pivotal role in inflammatory and immune responses [9]. Recent reports indicate that there is a significant interest for the identification of small-molecule antagonists of CXCR3 [10,11]. 4-*N*-Aryl-[1,4] diazepane ureas were found to constitute a promising series of functional antagonists of CXCR3 that could be developed into new therapeutic agents for the treatment of inflammatory disorders such us rheumatoid arthritis, inflammatory bowel disease, multiple sclerosis and diabetes [12].

In the past, several attempts have been made to build QSAR models in the general field of chemokine antagonists such as CCR5 [13,14], CXCR2 [15] and CXCR4 [16]. After a systematic literature search [17], we are confident that this paper presents the first QSAR study concerning small-molecule antagonists of CXCR3.

In particular, a series of 4-*N*-aryl-[1,4] diazepane ureas [12], recently discovered CXCR3 receptor antagonists, were studied in this work. Sixty-two physicochemical and topological descriptors were examined in terms of their efficacy to determine and predict the biological activity of the investigated derivatives. The descriptors were calculated using Topix [18] and Chem3D [19]. Among them, the most statistically significant descriptors were selected using the elimination selection-stepwise regression (ES-SWR) variable selection method. The result of this study was the development of a new linear QSAR model containing five variables. The proposed methodology was validated

* Corresponding authors. Biomedical Sciences Research Center "Alexander Fleming", 34 Fleming Street, Vari 16672, Greece. Tel.: +30 210 8979097; fax: +30 210 8979031.
E-mail addresses: afantitis@fleming.gr (A. Afantitis), kollias@fleming.gr (G. Kollias).

0223-5234/$ - see front matter © 2008 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2008.05.028

using several strategies: cross-validation, *Y*-randomization and external validation using division of the entire data set into training and test sets. Furthermore, the domain of applicability which indicates the area of reliable predictions was defined.

2. Materials and methods

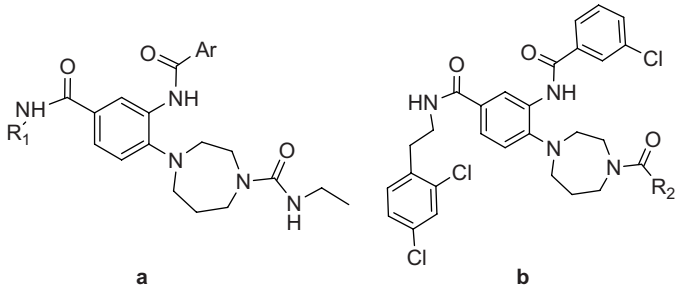
2.1. Data set

In this QSAR study, biological and chemical data from 32 diazepane ureas were used, which have been presented in the work of Cole et al. [12] (Table 1). In order to model and predict the biological effect of the specific compounds as functional antagonists of the chemokine receptor CXCR3, 62 physicochemical constants, topological and structural descriptors (Table 2) were considered as possible input candidates to the model. All the descriptors were calculated using Chem3D and Topix. Before the calculation of the descriptors, the structures 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 [19].

2.2. Separation into training and validation sets

The separation of the data set into training and validation sets was performed according to the popular Kennard and

Table 1
Data set and model predictions using Eq. (15)



Id	R_1	Ar	R_2	log (1/IC_{50})[b] (observed)	Training data[b] log (1/IC_{50}) (predicted); $R^2 = 0.82$;$R^2_{LOO} = 0.71$	Validation data[b] log (1/IC_{50}) (predicted) $R^2_{pred} = 0.75$	Leverages (limit = 0.60)
1a	2,4-Cl_2Ph	Ph	–	1.15	0.93	–	0.19
2a[a]	2,4-Cl_2Ph	2-Cl Ph	–	0.15	–	0.17	0.19
3a	2,4-Cl_2Ph	3-Cl Ph	–	1.22	0.93	–	0.14
4a	2,4-Cl_2Ph	3-MeO Ph	–	0.77	0.83	–	0.12
5a	2,4-Cl_2Ph	3-CN Ph	–	1.10	0.96	–	0.29
6a	2,4-Cl_2Ph	2-F Ph	–	0.23	0.12	–	0.20
7a[a]	2,4-Cl_2Ph	3-F Ph	–	1.22	–	0.80	0.07
8a	2,4-Cl_2Ph	4-F Ph	–	0.70	0.95	–	0.13
9a	2,4-Cl_2Ph	3,4-F_2 Ph	–	0.54	0.39	–	0.23
10a	2,4-Cl_2Ph	3,5-F_2 Ph	–	0.62	0.66	–	0.21
11a	2,4-Cl_2Ph	2-Thiophene	–	0.89	0.56	–	0.20
12a	2,4-Cl_2Ph	4-Pyridyl	–	0.60	0.52	–	0.39
13a	2,4-Cl_2Ph	3-Thiophene	–	0.31	0.56	–	0.20
14a[a]	2,4-Cl_2Ph	2-Furan	–	0.19	–	0.47	0.19
15a	2-Cl Ph Et	3-Cl Ph	–	0.72	0.42	–	0.11
16a	3-Cl Ph Et	3-Cl Ph	–	0.52	0.57	–	0.12
17a[a]	4-Cl Ph Et	3-Cl Ph	–	0.92	–	0.73	0.09
18a	2-F Ph Et	3-Cl Ph	–	0.47	0.29	–	0.09
19a	3-F Ph Et	3-Cl Ph	–	0.47	0.44	–	0.15
20a	4-F Ph Et	3-Cl Ph	–	0.49	0.60	–	0.13
21a	Ph Et	3-Cl Ph	–	0.29	0.61	–	0.57
22a	*c*Pr Ph Et	3-Cl Ph	–	0.62	0.31	–	0.27
23a	3,4-$(CH_3O)_2$PhEt	3-Cl Ph	–	0.38	0.50	–	0.50
24a	Bn	3-Cl Ph	–	−0.10	0.16	–	0.24
25a[a]	4-Cl Bn	3-Cl Ph	–	0.60	–	0.53	0.13
26a	^{i}Pr	3-Cl Ph	–	−0.33	−0.29	–	0.45
27a	*c*Pr Me	3-Cl Ph	–	−0.33	−0.29	–	0.34
28b[a]	–	–	$-NH^nPr$	1.22	–	1.26	0.19
29b	–	–	$-NH^iPr$	1.22	1.16	–	0.22
30b	–	–	$-NH^nBu$	1.30	1.51	–	0.35
31b	–	–	–NHMe	0.10	0.53	–	0.19
32b[a]	–	–	$-NMe_2$	0.05	–	0.42	0.37

[a] Validation set.
[b] IC_{50} in μM.

Table 2
Physicochemical constants, topological and structural descriptors

ID	Description	Notation	ID	Description	Notation
1	Molar refractivity	MR	2	Diameter	Diam
3	Partition coefficient (octanol–water)	ClogP	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 clusters	NClusters
43	Number of rings	NRings	44	Wiener Dim	Wiener Dim
45	Bertz	Bertz	46	AtomCompMean	AtomCompMean
47	AtomCompTot	AtomCompTot	48	Zagreb1	Zagreb1
49	Zagreb2	Zagreb2	50	Quadratic	Quadr
51	ScHultz	ScHultz	52	Kappa1	^{1}k
53	Kappa3	^{3}k	54	Kappa2	^{2}k
55	Wiener distance	WienerDistCode	56	Wiener information	InfWiener
57	DistEqMean	DistEqMean	58	DistEqTotal	DistEqTotal
59	InfMagnitDistTot	InfMagnitDistTot	60	Polarity	Polarity
61	Gordon	Gordon	62	Randic information 8	ChiInf8

Stones algorithm [20]. The algorithm starts by finding two samples that are 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 data set and placed into the calibration data set. This procedure is repeated until the desired number of samples has been reached in the validation data set. A commonly used ratio of training to validation objects, which is also adopted in this work, is 80%:20% [21]. 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 the test samples all fall inside the measured region. The Kennard and Stones algorithm has been applied with great success in many recent QSAR studies [22–27] and it has been highlighted as one of the best ways to build training and test sets [28].

2.3. Multiple linear regression (MLR) model development-variable selection

The first objective was to determine the optimum set of variables that produces the most significant linear QSAR models linking and interpreting the chemical structure of the small molecules with their functional activity. ES-SWR algorithm was used on the training data set to select the most appropriate descriptors. ES-SWR is a popular stepwise technique [29] that combines the advantages of both Forward Selection (FS-SWR) and Backward Elimination (BE-SWR). Forward Selection is computationally efficient for the generation of nested subsets of variables. On the other hand Backward Selection eliminates the most appropriate variable, so that the remaining variables perform best [30].

2.4. Cross-validation technique

Cross-validation is a popular technique used to explore the predictive ability of statistical models. Assuming that a training data set consisting of n available compounds, is available, a number of modified data sets are created by deleting in each case one or a small group (leave-some-out) of objects [29]. For each data set, an input–output model is developed, based on the utilized modeling technique. The model is evaluated by measuring its accuracy in predicting the responses of the remaining data (the ones that have not been utilized in the development of the model). In particular, the leave-one-out (LOO) and the leave-five-out (L5O) procedures were utilized in this study, which produce a number of models, by deleting one or five objects, respectively, from the training set. The maximum number of models produced by the LOO procedure is equal to the number of available examples n, while for the L5O procedure the maximum number of models is equal to $n!/5!(n-5)!$. Prediction error sum of squares (PRESS) is

a standard index to measure the accuracy of a modeling method using the LOO cross-validation technique [29]. Based on the PRESS statistic and the summation of squares of deviations of the experimental values from their mean (SSY), the squared correlation coefficient of predictions using the LOO method R^2_{LOO}, the standard error of predictions (SDEP) and the S_{PRESS} statistic can be easily calculated. The formulae used to calculate the aforementioned statistics are presented below:

$$R^2_{\mathrm{LOO}} = 1 - \frac{\mathrm{PRESS}}{\mathrm{SSY}} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \tilde{y}_{i,\mathrm{LOO}}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}_{\mathrm{tr}}\right)^2} \quad (1)$$

$$\mathrm{SDEP} = \sqrt{\frac{\mathrm{PRESS}}{n}} \quad (2)$$

$$S_{\mathrm{PRESS}} = \sqrt{\frac{\mathrm{PRESS}}{(n-k-1)}} \quad (3)$$

where $\overline{y}_{\mathrm{tr}}$ is the averaged value of the dependent variable for the training set, and y_i, $\tilde{y}_{i,\mathrm{LOO}}$, $i = 1, \ldots, n$ are the measured and predicted values of the dependent variable over the available training set. The squared correlation coefficients for the L50 cross-validation method R^2_{L5O} can be calculated in a similar manner. In particular, Eq. (1) is used to compute R^2_{L5O}, where the summations run over the predictions of all models that are produced by deleting five objects from the training set.

2.5. Quality of fit and predictive ability of a QSAR model

The first indication of the success on a QSAR model is to measure the quality of fit on the available training data. The most common objective criteria [29] used for this purpose are the squared correlation coefficient R^2, the root mean squared (RMS) error statistic and the F-value, and, which are defined next:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}_{\mathrm{tr}}\right)^2} \quad (4)$$

$$\mathrm{RMS} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2}{(n-k-1)}} \quad (5)$$

$$F = \frac{\left(R^2/k\right)}{\left(\left(1-R^2\right)/(n-k-1)\right)} \quad (6)$$

In the above equations, k is the number of independent variables in the model and $\widehat{y}_i$, $i = 1, \ldots, n$ are the values calculated by the QSAR model for the dependent variable. We should note here that $\widehat{y}_i$ is the value calculated by the QSAR model for the dependent variable corresponding to object i when this compound has been included in the training data set, whereas $\tilde{y}_{i,\mathrm{LOO}}$ is the prediction of the model that has not utilized compound i throughout the model development procedure.

According to Tropsha et al. [31] the predictive ability of a QSAR model should be tested on an external set of data that has not been taken into account during the process of developing the model. In particular, the following statistical indices have been proposed [31,32] to assess the predictive power of QSAR models, besides the popular squared correlation coefficient R^2_{pred}:

$$R^2_{\mathrm{ext}} = 1 - \frac{\sum_{i=1}^{\mathrm{ntest}}\left(y_i - \tilde{y}_i\right)^2}{\sum_{i=1}^{\mathrm{ntest}}\left(y_i - \overline{y}_{\mathrm{tr}}\right)^2} \quad (7)$$

$$k = \frac{\sum_{i=1}^{\mathrm{ntest}} y_i \tilde{y}_i}{\sum_{i=1}^{\mathrm{ntest}} \tilde{y}_i^{\,2}} \quad (8)$$

$$R^2_{\mathrm{o}} = 1 - \frac{\sum_{i=1}^{\mathrm{ntest}}\left(\tilde{y}_i - y_i^{\mathrm{ro}}\right)^2}{\sum_{i=1}^{\mathrm{ntest}}\left(\tilde{y}_i - \overline{\tilde{y}}\right)^2}, \text{ where } y_i^{\mathrm{ro}} = k\tilde{y}_i, \;\; i = 1, \ldots, \mathrm{ntest} \quad (9)$$

In the above equation ntest is the number of compounds that constitute the validation data set, $\overline{y}_{\mathrm{tr}}$ is the averaged value of the dependent variable for the training set, y_i, $\tilde{y}$, $i = 1, \ldots,$ ntest are the measured values and the QSAR model predictions of the dependent variable over the available validation set and $\overline{\tilde{y}}$ is the average over all $\tilde{y}$, $i = 1, \ldots,$ ntest.

Tropsha et al. [31,32] considered a QSAR model to be predictive, if the following conditions are satisfied:

$$R^2_{\mathrm{ext}} > 0.5 \quad (10)$$

$$R^2_{\mathrm{pred}} > 0.6 \quad (11)$$

$$\frac{\left(R^2_{\mathrm{pred}} - R^2_{\mathrm{o}}\right)}{R^2_{\mathrm{pred}}} < 0.1 \quad (12)$$

$$0.85 \le k \le 1.15 \quad (13)$$

2.6. Defining model applicability domain

In order for a QSAR model to be used for screening new compounds, its domain of application [31,30] must be defined and predictions for only those compounds that fall into this domain may be considered reliable. *Extent of extrapolation* [32] is one simple approach to define the applicability of the domain. It is based on the calculation of the leverage h_i [32] for each chemical, where the QSAR model is used to predict its activity:

$$h_i = x_i^{\mathrm{T}}\left(X^{\mathrm{T}}X\right)x_i \quad (14)$$

In Eq. (14) 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 not be reliable.

2.7. Y-randomization test

This technique ensures the robustness of a QSAR model [33]. The dependent variable vector is randomly shuffled and a new QSAR model is developed using the original independent variable matrix [23,34–36]. The new QSAR models (after several repetitions) are expected to have low R^2 and R^2_{LOO} values. If the opposite happens then an acceptable QSAR model cannot be obtained for the specific modeling method and data.

3. Results and discussion

First, the data set of 32 derivatives was partitioned into a training set of 25 compounds, and a validation set of 7 compounds according to the Kennard and Stones algorithm [20] using a 80%:20% ratio as mentioned before [21]. The validation examples are represented by table footnote "a" in Table 1. The algorithm was applied on the complete database consisting of all 62 available descriptors (please see Table 2). The validation data were not involved by any means in the process of selecting the most appropriate descriptors or in the development of the QSAR model. They were considered as a completely unknown external set of data, which was used only to test the accuracy of the produced model. The MLR QSAR model was thus developed by applying the ES-SWR algorithm on the set of training data. The result was the following five-variable equation:

$$\log(1/\mathrm{IC}_{50}) = 3.99 - 2.58\,\mathrm{ChiInf0} - 2.35\,\mathrm{ChiInf8} - 8.85\times10^{-3}\,\mathrm{AtomCompTot} + 7.78\times10^{-1}\,{}^3k + 1.94\times10^{-1}\,\mathrm{ClogP} \quad (15)$$

$$R^2 = 0.82 \quad \mathrm{RMS} = 0.21 \quad F = 16.94 \quad R^2_{LOO} = 0.71 \quad \mathrm{SDEP} = 0.23 \quad S_{PRESS} = 0.27 \quad n = 25$$

Table 3 presents the correlation matrix, where it is clear that the five selected descriptors are not highly correlated. Another important observation is that the ratio of the objects in the training set to the number of descriptors is 5:1, which is the case in many QSAR [37].

The five input variables in the QSAR model are measured in different units of measurements and the respective coefficients are of different orders of magnitude. In order to examine the importance of each descriptor and answer the question which of the independent variables have a greater effect on the dependent variable in the multiple regression analysis, the standardized regression coefficients were also calculated.

Table 3
Correlation matrix for the five selected descriptors

	ChiInf0	ChiInf8	AtomCompTot	^3k	*C* Log *P*
ChiInf0	1				
ChiInf8	−0.12	1			
AtomCompTot	−0.10	0.16	1		
Kappa3	0.21	0.34	0.71	1	
ClogP	0.32	0.19	0.25	0.66	1

This calculation is performed by applying the multiple regression methodology on the standardized values of the independent and dependent variables, i.e. on the values that are obtained after subtracting the mean and dividing by the standard deviation for each variable [29]. The standardized regression coefficients, then, represent the change in a dependent variable that results from a change of one standard deviation in an independent variable. The standardized regression coefficients are presented in the following QSAR model:

$$\log(1/\mathrm{IC}_{50}) = -0.25\,\mathrm{ChiInf0} - 0.50\,\mathrm{ChiInf8} - 0.31\,\mathrm{AtomCompTot} + 0.93\,{}^3k + 0.34\,\mathrm{ClogP} \quad (16)$$

It is clear that the standardized regression coefficients for all input descriptors are of the same scale. We can conclude that all descriptors are significant and are of similar importance for the investigated activity.

The model was quite stable to the inclusion–exclusion of compounds measured by the LOO and L5O cross-validation procedures. This is indicated by the following statistics:

$$R^2_{LOO} = 0.71$$

$$R^2_{L5O} = 0.69$$

R^2_{LOO} and R^2_{L5O} are calculated using only the 25 training examples. Calculation of the R^2_{LOO} statistic was performed using all 25 models that are produced by excluding one compound each time from the training examples, while calculation of the R^2_{L5O} statistic was based on 1000 random exclusions of five-member groups of examples.

The model (Eq. (15)) also passed Tropsha's [31,32] recommended tests for predictive ability (Eq. (10–13)):

$$R^2_{ext} = 0.72 > 0.5$$

$$R^2_{pred} = 0.75 > 0.6$$

$$\frac{\left(R^2_{pred} - R^2_{o}\right)}{R^2_{pred}} = -0.30 < 0.1$$

$$k = 1.05 \approx 1$$

The model was further validated by applying the *Y*-randomization test. In particular, 10 random shuffles of the *Y*-vector gave R^2 and R^2_{LOO} values in the ranges of 0.1–0.30 and 0.05–0.25, respectively. The low R^2 and R^2_{LOO} values that were obtained show that the good results in our original model are not due to a chance correlation or structural dependency of the training set.

Remark: according to some researchers [33], in each cycle of the *Y*-test, the entire variable selection procedure should be carried out on the scrambled data. This modified *Y*-randomization test was also performed in our model. The obtained R^2 and R^2_{LOO} values were lower than 0.22 for all random shuffles of the *Y*-vector that were examined.

The above results illustrate that the linear MLR technique combined with a successful variable selection procedure is adequate to generate a successful QSAR model for modeling and predicting the functional antagonist of CXCR3 by 4-*N*-aryl-[1,4] diazepane ureas.

It needs to be emphasized, however, that no matter how robust, significant and validated a QSAR model may be, it cannot be expected to predict reliably the modeled activity for the entire universe of chemicals [31,32]. The domain of applicability of the model was defined using the extent of extrapolation method [27,36]. According to this method, we consider as reliable only the predictions of the compounds, whose leverages lie within the domain of applicability. In Table 1 all leverages for the training and test sets are presented. The warning leverage limit is 0.60 and as it can be concluded from the leverage values in Table 1, the predictions of the QSAR model for all the compounds (both the training and test sets) are considered reliable.

The chemical meaning of the five descriptors used in the produced QSAR model is briefly described next.

Kier shape descriptors (kappa indices) derived from the counts of atoms and bonds depict a molecule as being related to the extremes of linear and maximally branched structures. Kappa indices encode information such as cyclicity, spatial density, symmetry and degree of centralization separation in branching. [35]. ${}^3k$ is the third order shape attribute which is described by the counts of three contiguous bonds ${}^3P$. For third order attribute, ${}^3P_{\max}$ is the three-bond paths in the twin star structure and ${}^3P_{\min}$ is the number of three-bond paths in the linear graph (Fig. 1).

The equation for calculating the ${}^3k$ index is as follows:

$$ {}^3k = \frac{4\,{}^3P_{\max}\,{}^3P_{\min}}{({}^3P_i)^2} \tag{17} $$

Information indices (Chinf0, Chinf8 and AtomCompTot) encode information on the adjacency and distance of atoms and the atomic composition in the molecular structure [29].

Topological information indices (Chinf0, Chinf8) are graph theoretical invariants that view the molecular graph as a source of different probability distributions to which the information theory is applied [29]. These indices have several advantages such as unique representation of the compound and high discriminating power (isomer discrimination). In a recent work, topological information descriptors were used with great success [39]. Information connectivity (Chinf0, Chinf8) indices are based on the partition of the edges in the graph according to the equivalence and the magnitudes of their edge connectivity values [29,38,40].

Let a given system I having n elements be regarded according to a certain equivalence relation, into k equivalence classes with cardinalities n_i where $n = \sum_{i=1}^{k} n_i$.

The information content of a system I with n elements is defined by the following equation (Eq. (18)), where the binary logarithm is used for measuring the information contents in bits.

$$ I = n \log_2 n - \sum_{i=1}^{k} n_i \log_2 n_i \tag{18} $$

Total information content on atomic composition AtomCompTot (I_{AC}) [29] is calculated from the complete molecular formula, hydrogen included, using the following equation:

$$ I_{AC} = A^{h} \log_2 A^{h} - \sum_{g} A_g \log_2 A_g \tag{19} $$

where A^{h} is the total number of atoms (hydrogen included) and A_g is the number of equal-type atoms in the gth equivalence class.

Lipophilicity is known to be important for absorption, permeability, and in vivo distribution of organic compounds [41] and has been used as a physicochemical in QSAR studies with great success [42,43].

According to the produced QSAR model (Eq. (15)) high values of the Kier shape descriptor 3 (${}^3k$) and lipophilicity (ClogP) contribute positively to the activity. Thus, an improvement in the activity is expected by designing small molecules that include the fragments depicted in Fig. 2, which encode information about the branching of acyclic structures. ${}^3k$ encodes structural features related to the central positioning of branching in a molecule. Moreover, the introduction of lipophilicity groups into diazepane urea's core will also affect the activity positively.

On the other hand, the information indices (Chinf0, Chinf8 and AtomCompTot), which encode information about the adjacency and distance of atoms and the atomic composition in the molecular structure [29], contribute negatively to the activity. The topological information indices (ChiInf0 and ChiInf8) measure the lack of homogeneity or the diversity of a molecular structure, and thus, they receive higher values for rather asymmetric structures [29]. Due to the negative contribution on the activity under study, we need to design compounds, for which these descriptors receive low values (i.e. design molecules with high levels of homogeneity). AtomCompTot is an information index [29] of the elemental composition of the molecule that takes into account the molecular diversity in terms of different atom types. In order to design molecules

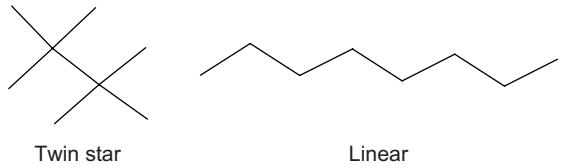


Fig. 1. Graphs of ${}^3P_{\max}$ and ${}^3P_{\min}$.

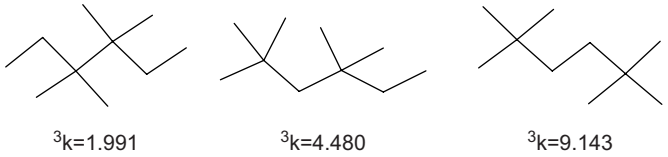


Fig. 2. ${}^3k$ values and corresponding shapes.

with low values for this descriptor, the proposed structures should not contain many different atom types.

The proposed method, due to the high predictive ability [31], and simplicity [44,45] could be a useful aid to the costly and time-consuming experiments for determining the CXCR3 functional antagonism effect of the diazepane ureas. A virtual screening procedure [46,47] could be based on the proposed QSAR model. The design of novel active molecules by the insertion, deletion or modification of substituents on different sites of the molecule and at different positions could be guided by the proposed model [4,14,27] .The method can also be used to screen existing databases or virtual combinations in order to identify derivatives with desired activity. In this case, the applicability domain will serve as a valuable tool to filter out ''dissimilar'' combinations.

4. Conclusion

The successful results of this study led to the conclusion that activity of small-molecule antagonists of CXCR3 can be successfully modeled with physicochemical constants and structural descriptors. The validation procedures (cross-validation, separation of data into independent training and validation sets, *Y*-randomization) illustrated the accuracy and robustness of the produced QSAR model not only by calculating its fitness on sets of training data, but also by testing the predictive ability of the model. The molecular descriptors used in QSAR encode information about the structure, branching, electronic effects, chains and rings of the modules and thus implicitly account for cooperative effects between functional groups. The proposed QSAR model aims at helping the researchers to design novel chemistry driven molecules with desired biological activity.

Acknowledgments

This work was supported by funding under the Sixth Research Framework Programme of the European Union, Project MUGEN (MUGEN LSHG-CT-2005-005203).

Appendix. Supplementary data

Supplementary data associated with this article can be found in the online version, at doi:10.1016/j.ejmech.2008.05.028.

References

[1] K.V. Camarda, C.D. Maranas, Ind. Eng. Chem. Res. 38 (1999) 1884–1892.
[2] D.K. Agrafiotis, D. Bandyopadhyay, J.K. Wegner, H. van Vlijmen, J. Chem. Inf. Model. 47 (2007) 1279–1293.
[3] P.R. Duchowicz, A. Talevi, C. Bellera, L.E. Bruno-Blanch, E.A. Castro, Bioorg. Med. Chem. 15 (2007) 3711–3719.
[4] G. Melagraki, A. Afantitis, H. Sarimveis, O. Igglessi-Markopoulou, A. Alexandridis, Mol. Divers. 10 (2006) 213–221.
[5] M.S. Castilho, R.V.C. Guido, A.D. Andricopulo, Bioorg. Med. Chem. 15 (2007) 6242–6252.
[6] M. Jalali-Heravi, A. Kyani, Eur. J. Med. Chem. 42 (2007) 649–659.
[7] S. Deswal, N. Roy, Eur. J. Med. Chem. 42 (2007) 463–470.
[8] B. Xia, W. Ma, B. Zheng, X. Zhang, B. Fan, Eur. J. Med. Chem. (2007). doi:10.1016/j.ejmech.2007.09.004.
[9] M. Loetscher, B. Gerber, P. Loetscher, S.A. Jones, L. Piali, I. Clark-Lewis, M. Baggioloni, B.J. Moser, Exp. Med. 184 (1996) 963–969.
[10] R.J. Watson, D.R. Allen, H.L. Birch, G.A. Chapman, F.C. Galvin, L.A. Jopling, R.L. Knight, D. Meier, K. Oliver, J.W.G. Meissner, D.A. Owen, E.J. Thomas, N. Tremayne, S.C. Williams, Bioorg. Med. Chem. Lett. 18 (2007) 147–151.
[11] D.R. Allen, A. Bolt, G.A. Chapman, R.L. Knight, J.W.G. Meissner, D.A. Owen, R.J. Watson, Bioorg. Med. Chem. Lett. 17 (2007) 697–701.
[12] A.G. Cole, I.L. Stroke, M.-R. Brescia, S. Simhadri, J.J. Zhang, Z. Hussain, M. Snider, C. Haskell, S. Ribeiro, K.C. Appell, I. Henderson, M.L. Webb, Bioorg. Med. Chem. Lett. 16 (2006) 200–203.
[13] Y.D. Aher, A. Agrawal, P.V. Bharatam, P. Garg, J. Mol. Model. 13 (2007) 519–529.
[14] A. Afantitis, G. Melagraki, H. Sarimveis, P.A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou, J. Comput. Aided Mol. Des. 20 (2006) 83–95.
[15] A.I. Khlebnikov, I.A. Schepetkin, M.T. Quinn, Bioorg. Med. Chem. 14 (2006) 352–365.
[16] J.B. Bhonsle, Z.-X. Wang, H. Tamamura, N. Fujii, S.C. Peiper, J.O. Trent, QSAR Comb. Sci. 24 (2005) 620–630.
[17] SciFinder 2007 & Scopus (accessed 10 March 2008).
[18] Topix.<www.lohninger.com/topix.html>.
[19] Chem3D.<www.cambridgesoft.com>.
[20] R.W. Kennard, L.A. Stone, Technometrics 11 (1969) 137–148.
[21] A.K. Chakraborti, B. Gopalakrishnan, M. Elizabeth Sobhia, A. Malde, Eur. J. Med. Chem. 38 (2003) 975–982.
[22] G. Melagraki, A. Afantitis, H. Sarimveis, P.A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou, J. Comput. Aided Mol. Des. 21 (2007) 251–267.
[23] A. Afantitis, G. Melagraki, H. Sarimveis, P.A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou, Bioorg. Med. Chem. 14 (2006) 6686–6694.
[24] A. Afantitis, G. Melagraki, H. Sarimveis, P.A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou, QSAR Comb. Sci. 25 (2006) 928–935.
[25] P. Ghosh, M. Thanadath, M.C. Bagchi, Mol. Divers. 10 (2006) 415–427.
[26] P. Fossa, L. Mosti, F. Bondavalli, S. Schenone, A. Ranise, C. Casolino, M. Forina, Bioorg. Med. Chem. 14 (2006) 1348–1363.
[27] G. Melagraki, A. Afantitis, H. Sarimveis, P.A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou, Bioorg. Med. Chem. 15 (2007) 7237–7247.
[28] W. Wu, B. Walczak, D.L. Massart, S. Heuerding, F. Erni, I.R. Last, K.A. Prebble, Chemometr. Intell. Lab. Syst. 33 (1996) 35–46.
[29] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, 2000.
[30] I. Guyon, A. Elisseeff, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[31] A. Tropsha, P. Gramatica, V.K. Gombar, QSAR Comb. Sci. 22 (2003) 69–77.
[32] A. Golbraikh, A. Tropsha, J. Mol. Graph. Model. 20 (2002) 269–276.
[33] K. Baumann, Trends Anal. Chem. 22 (2003) 395–406.
[34] C. Pramod, M. Nair, E. Sobhia, Eur. J. Med. Chem. 43 (2008) 293–299.
[35] M. Jalali-Heravi, M. Asadollahi-Baboli, P. Shahbazikhah, Eur. J. Med. Chem. 43 (2008) 548–556.
[36] A. Afantitis, G. Melagraki, H. Sarimveis, P.A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou, Mol. Divers. 10 (2006) 405–414.
[37] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, Anal. Chim. Acta 515 (2004) 199–208.
[38] J. Devillers, A.T. Balaban (Eds.), Topological Indices and Related Descriptors in QSAR and QSPR, Gordon and Breach Science Publishers, The Netherlands, 1999, pp. 455–489.
[39] A. Afantitis, G. Melagraki, H. Sarimveis, O. Igglessi-Markopoulou, C.T. Supuran, Bioorg. Med. Chem. 14 (2006) 1108–1114.
[40] E.V. Konstantinova, Electron. Notes Discrete Math. 21 (2005) 329–351.

[41] R. Mannhold, A. Petrauskas, QSAR Comb. Sci. 22 (2003) 466–475.
[42] Y. Zhou, L. Zhu, Y. Tang, D. Ye, Eur. J. Med. Chem. 42 (2007) 977–984.
[43] J. Matysiak, Eur. J. Med. Chem. 42 (2007) 940–947.
[44] M. Hewitt, M.T.D. Cronin, J.C. Madden, P.H. Rowe, C. Johnson, A. Obi, S.J. Enoch, J. Chem. Inf. Model. 47 (2007) 1460–1468.
[45] C.W. Yap, H. Li, Z.L. Ji, Y.Z. Chen, Mini Rev. Med. Chem. 11 (2007) 1097–1107.
[46] R.V.C. Guido, G. Oliva, A.D. Andricopulo, Curr. Med. Chem. 15 (2008) 37–46.
[47] I. Muegge, S. Oloff, Drug Discov. Today Technol. 3 (2006) 405–411.