Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.jfda-online.com

Special Invited Article

Amino substituted nitrogen heterocycle ureas as kinase insert domain containing receptor (KDR) inhibitors: Performance of structure–activity relationship approaches

Hayriye Yilmaz [a,b], *Natalia Sizochenko* [b,c], *Bakhtiyor Rasulev* [b], *Andrey Toropov* [d], *Yahya Guzel* [e], *Viktor Kuz'min* [c], *Danuta Leszczynska* [f], *Jerzy Leszczynski* [b,*]

[a] *Kayseri Vocational School, Biomedical Devices and Technologies, Erciyes University, 38039, Kayseri, Turkey*
[b] *Interdisciplinary Center for Nanotoxicity, Department of Chemistry and Biochemistry, Jackson State University, Jackson, MS, 39217, USA*
[c] *Odessa I.I. Mechnikov National University, Department of Chemistry, Dvoryanskaya Street, 2, 65082, Odessa, Ukraine*
[d] *Laboratory of Environmental Chemistry and Toxicology, IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, 20156, Via La Masa 19, Milano, Italy*
[e] *Department of Chemistry, Faculty of Science, Erciyes University, 38039, Kayseri, Turkey*
[f] *Department of Civil and Environmental Engineering, Jackson State University, Jackson, MS, 39217, USA*

ARTICLE INFO

Article history:
Available online 1 April 2015

Keywords:
amino-substituted nitrogen heterocyclic ureas
descriptors
KDR inhibitors
QSAR
SiRMS
SMILES

ABSTRACT

A quantitative structure–activity relationship (QSAR) study was performed on a set of amino-substituted nitrogen heterocyclic urea derivatives. Two novel approaches were applied: (1) the simplified molecular input-line entry systems (SMILES) based optimal descriptors approach; and (2) the fragment-based simplex representation of molecular structure (SiRMS) approach. Comparison with the classic scheme of building up the model and balance of correlation (BC) for optimal descriptors approach shows that the BC scheme provides more robust predictions than the classic scheme for the considered pIC_{50} of the heterocyclic urea derivatives. Comparison of the SMILES-based optimal descriptors and SiRMS approaches has confirmed good performance of both techniques in prediction of kinase insert domain containing receptor (KDR) inhibitory activity, expressed as a logarithm of inhibitory concentration (pIC50) of studied compounds.

* *Corresponding author*. Interdisciplinary Nanotoxicity Center, Department of Chemistry and Biochemistry, Jackson State University, 1400 J. R. Lynch Street, Post Office Box 17910, Jackson, MI 39217, USA.
E-mail address: jerzy@icnanotox.org (J. Leszczynski).
http://dx.doi.org/10.1016/j.jfda.2015.03.001

1. Introduction

The kinase insert domain containing receptor (KDR), alternatively referred to as VEGFR-2, is a receptor for vascular endothelial growth factors (VEGFs). It functions as a key regulator of angiogenesis, the process by which new capillaries are created from preexisting blood vessels [1]. Accordingly, interruption of VEGFR-2 signaling by small molecule inhibitors to VEGFR-2 kinase domain has been shown to be an attractive strategy in the treatment of cancer. In recent years, a novel series of amino-substituted nitrogen heterocyclic urea derivatives has been reported as being essential inhibitors against KDR [2].

Quantitative structure–activity relationship (QSAR) methods are widely applied nowadays to find mathematical relationships between the chemical structure of a compound and its biological activity [3–17]. This technique was utilized here, based on experimental data available, and calculated theoretical descriptors, to perform an inhibitory activity study [6,10,17]. In the present study, the predictive QSAR models were developed for a set of amino-substituted nitrogen heterocyclic ureas for which the molecular structure is represented by simplified molecular input-line entry systems (SMILES) applying new techniques, such as the SMILES-based such as the SMILES-based optimal descriptors approach implemented in CORrelations And Logic (CORAL) (http://www.insilico.eu/coral), and the simplex representation of molecular structure (SiRMS) approach [18].

2. Materials and methods

2.1. Dataset

For prediction of inhibitory binding affinities (pIC_{50}, i.e., logarithm of the 50% effective concentration) the data on 63 amino-substituted nitrogen heterocyclic ureas were collected from existing literature [19].

2.2. Computational details

2.2.1. CORAL approach

There are three options for the selection of optimal descriptors in CORAL: (1) graph based; (2) SMILES based; and (3) hybrid descriptors which are calculated using both graph and SMILES approaches [20–23]. There are two classes of graph invariants which are available in CORAL: vertices and Morgan vertices degrees. In the case of hydrogen-suppressed graphs (HSGs) and hydrogen-filled graphs, vertices are representations of the chemical elements, such as carbon, nitrogen, oxygen, etc. In the case of graphs of atomic orbitals, vertices represent electronic structures i.e. atomic orbitals such as $1s^1$, $2s^2$, $2p^5$, $3d^{10}$, etc. [24].

The optimal graph-based descriptor based on so-called correlation weights (DCW) is calculated as the following:

$$^{Graph}(\text{Threshold}, N_{epoch}) = \sum CW(A_k) + a\sum CW(^0EC_k) + \beta\sum CW(^1EC_k) + \gamma CW(^2EC_k) + \delta CW(^3EC_k) \quad \text{Equation 1}$$

Three topological invariants of the molecular graphs were involved in current study: vertex degree (EC0); extended connectivity of first order (EC1); and extended connectivity of second order (EC2) [25].

The optimal SMILES-based descriptor based on correlation weights:

$$^{SMILES}(\text{Threshold}, N_{epoch}) = a\sum CW(S_k) + \beta\sum CW(SS_k) + \gamma\sum CW(SSS_k) + \delta CW(\text{PAIR}) + x\,CW(\text{NOSP}) + y\,CW(\text{HALO}) + z\,CW(\text{BOND}) \quad \text{Equation 2}$$

S_k, SS_k, and SSS_k are representations of molecular fragments, for example if SMILES = Clc1ccccc1 then s_k = (Cl, c, 1, c, c, c, c, 1); ss_k = (Clc, c1, cc, cc, cc, cc, c1); sss_k = (Clc1, c1c, ccc, ccc, ccc, ccc, cc1). *PAIR, NOSP, HALO,* and *BOND* are global SMILES attributes which are calculated with SMILES. These global attributes provide the possibility of carrying out an additional discrimination of substances into separated classes: for example nitrogen, oxygen, sulphur, and phosphorus (NOSP); fluorine, chlorine, and bromine (HALO) [24]. The BOND attribute is related to presence/absence of three categories of chemical bonds: double, triple, and stereospecific. The coefficients *a*, β, γ, x, y, and z can be either 1 or 0. One (1) indicates that the SMILES attribute is involved in the calculation of the descriptor of correlation weights (DCW) (Threshold) and zero (0) indicates that the SMILES attribute is not involved. Combinations of values of different attributes provide the possibility of defining various versions of SMILES based optimal descriptors [20].

CORAL software can be also used to build up a hybrid model which is calculated with SMILES-based and GRAPH-based descriptors:

$$^{Hybrids}(\text{Threshold}, N_{epoch}) = {}^{Graph}(\text{Threshold}, N_{epoch}) + {}^{SMILES}(\text{Threshold}, N_{epoch}) \quad \text{Equation 3}$$

The graph- and SMILES-based models are mathematical functions of the threshold and the number of N_{epoch} of the Monte Carlo optimization. The most predictive combination of T and N_{epoch} values for a split of data can be found by analyzing results of the calculations for several different splits of data in the training and test sets.

2.2.2. SiRMS approach

In addition to the above mentioned approaches, the SiRMS technique [18] was also applied to calculate fragmentary 2D descriptors (fragments of the size 2–5). In the framework of SiRMS, any molecule can be represented as a system of different simplexes (fragments of fixed composition and topology). In previous studies this method provided good results for solving different "structure–activity" problems [26–29].

In the current study a 2D level of molecule representation was utilized to generate simplex fragments. During the first step, the connectivity of atoms in simplex, atom type, and bond nature were considered. For each property the range is created with four to seven intervals. In this study all atoms were divided into groups corresponding to their atomic refraction (A < 1.5 < B < 3 < C < 8 < D), partial charges

Table 1 – Statistical quality of models developed by the CORAL approach.

Trshd	N_{act}	Probe	Training set				Calibration set				Test set				
			n_t	r_t^2	s_t	F_t	n_c	r_c^2	s_c	F_c	n_v	r_v^2	s_v	F_v	R_m^2
Split 1 Balance of correlations															
0	93	1	39	0.8506	0.313	211	13	0.9815	0.531	585	11	0.7537	0.497	28	0.7369
0	**93**	**2**	**39**	**0.8522**	**0.312**	**213**	**13**	**0.9851**	**0.510**	**728**	**11**	**0.7873**	**0.465**	**33**	**0.7834**
0	93	3	39	0.8496	0.314	209	13	0.9873	0.513	852	11	0.7502	0.500	27	0.7468
0				0.8508	0.313	211		0.9846	0.518	722		0.7637	0.487	29	0.7557
1	90	1	39	0.8482	0.316	207	13	0.9887	0.506	962	11	0.7413	0.513	26	0.7320
1	90	2	39	0.8460	0.318	203	13	0.9850	0.517	723	11	0.7526	0.505	27	0.7408
1	90	3	39	0.8505	0.313	210	13	0.9871	0.516	840	11	0.7047	0.548	21	0.6737
1				0.8482	0.316	207		0.9869	0.513	842		0.7329	0.522	25	0.7155
2	73	1	39	0.8329	0.331	184	13	0.9816	0.580	587	11	0.7218	0.527	23	0.7212
2	73	2	39	0.8403	0.324	195	13	0.9762	0.554	450	11	0.7309	0.519	24	0.7065
2	73	3	39	0.8294	0.335	180	13	0.9839	0.560	674	11	0.7245	0.528	24	0.6873
2				0.8342	0.330	186		0.9806	0.565	570		0.7257	0.524	24	0.7050
Split 1 Classic scheme															
0	93	1	52	0.8648	0.295	320					11	0.7990	0.485	36	0.6607
0	93	2	52	0.8647	0.295	319					11	0.7137	0.591	22	0.5350
0	93	3	52	0.8630	0.297	315					11	0.7789	0.515	32	0.6551
0				0.8642	0.295	318						0.7639	0.530	30	0.6169
1	90	1	52	0.8654	0.294	321					11	0.7616	0.535	29	0.6273
1	90	2	52	0.8652	0.294	321					11	0.7855	0.510	33	0.6356
1	90	3	52	0.8668	0.293	325					11	0.8001	0.531	36	0.6684
1				0.8658	0.294	323						0.7824	0.525	33	0.6438
2	73	1	52	0.8588	0.301	304					11	0.7484	0.525	27	0.6310
2	73	2	52	0.8564	0.304	298					11	0.7655	0.497	29	0.6812
2	73	3	52	0.8544	0.306	293					11	0.7564	0.504	28	0.6847
2				0.8565	0.304	299						0.7568	0.508	28	0.6656
Split 2 Balance of correlations															
0	29	1	42	0.8023	0.395	162	11	0.8870	0.439	71	10	0.7752	0.400	28	0.5824
0	29	2	42	0.8036	0.394	164	11	0.8848	0.425	69	10	0.7629	0.417	26	0.5875
0	29	3	42	0.8024	0.395	162	11	0.8879	0.438	71	10	0.5886	0.560	11	0.4300
0				0.8028	0.395	163		0.8866	0.434	70		0.7089	0.459	22	0.5333
1	29	1	42	0.8027	0.395	163	11	0.8857	0.417	70	10	0.6891	0.479	18	0.5334
1	29	2	42	0.8002	0.397	160	11	0.8868	0.439	71	10	0.6689	0.496	16	0.5318
1	29	3	42	0.8025	0.395	163	11	0.8872	0.446	71	10	0.7280	0.438	21	0.5773
1				0.8018	0.396	162		0.8866	0.434	70		0.6953	0.471	18	0.5475
2	28	1	42	0.6852	0.499	87	11	0.8873	0.389	71	10	0.7887	0.397	30	0.5450
2	28	2	42	0.6867	0.497	88	11	0.8870	0.389	71	10	0.6295	0.532	14	0.3921
2	28	3	42	0.6838	0.500	87	11	0.8898	0.404	73	10	0.7627	0.409	26	0.5067
2				0.6852	0.499	87		0.8881	0.394	71		0.7270	0.446	23	0.4813
Split 2 Classic scheme															
0	29	1	53	0.8078	0.383	214					10	0.7227	0.500	21	0.5357
0	29	2	53	0.8057	0.385	211					10	0.7701	0.454	27	0.5577
0	29	3	53	0.8067	0.384	213					10	0.5627	0.633	10	0.3850
0				0.8067	0.384	213						0.6852	0.529	19	0.4928
1	29	1	53	0.8081	0.382	215					10	0.7612	0.458	25	0.5240
1	29	2	53	0.8075	0.383	214					10	0.7213	0.498	21	0.5341
1	29	3	53	0.8078	0.383	214					10	0.7077	0.511	19	0.5139
1				0.8078	0.383	214						0.7301	0.489	22	0.5240
2	28	1	53	0.7094	0.470	125					10	0.7628	0.474	26	0.5223
2	28	2	53	0.7123	0.468	126					10	0.7498	0.474	24	0.4727
2	28	3	53	0.7124	0.468	126					10	0.6805	0.540	17	0.4303
2				0.7114	0.469	126						0.7311	0.496	22	0.4751
Split 3 Balance of correlations															
0	31	1	40	0.7755	0.372	131	13	0.9646	0.348	300	10	0.5978	0.844	12	0.5730
0	31	2	40	0.7762	0.371	132	13	0.9628	0.355	285	10	0.5654	0.896	10	0.5378
0	31	3	40	0.7734	0.373	130	13	0.9662	0.357	314	10	0.5860	0.857	11	0.5695
0				0.7750	0.372	131		0.9645	0.354	300		0.5831	0.866	11	0.5601
1	31	1	40	0.7775	0.370	133	13	0.9641	0.361	295	10	0.5876	0.862	11	0.5638
1	31	2	40	0.7737	0.373	130	13	0.9659	0.355	312	10	0.5792	0.857	11	0.5701
1	31	3	40	0.7758	0.371	132	13	0.9603	0.333	266	10	0.5612	0.918	10	0.5226
1				0.7757	0.371	131		0.9634	0.350	291		0.5760	0.879	11	0.5522
2	28	1	40	0.6032	0.494	58	13	0.9515	0.273	216	10	0.6680	0.910	16	0.4494
2	28	2	40	0.6039	0.494	58	13	0.9529	0.269	223	10	0.6675	0.910	16	0.4503

Table 1 – (continued)

Trshd	N_{act}	Probe	Training set				Calibration set				Test set				
			n_t	r_t^2	s_t	F_t	n_c	r_c^2	s_c	F_c	n_υ	r_υ^2	s_υ	F_υ	R_m^2
2	28	3	40	0.6069	0.492	59	13	0.9511	0.264	214	10	0.6727	0.896	16	0.4597
2				0.6047	0.493	58		0.9518	0.269	217		0.6694	0.905	16	0.4532
Split 3 Classic scheme															
0	29	1	53	0.8125	0.338	221					10	0.6292	0.777	14	0.6130
0	29	2	53	0.8131	0.338	222					10	0.6174	0.794	13	0.5942
0	29	3	53	0.8119	0.339	220					10	0.6407	0.756	14	0.6104
0				0.8125	0.338	221						0.6291	0.776	14	0.6058
1	29	1	53	0.8125	0.338	221					10	0.6233	0.789	13	0.6047
1	29	2	53	0.8130	0.338	222					10	0.6139	0.795	13	0.5857
1	29	3	53	0.8120	0.339	220					10	0.6307	0.769	14	0.6062
1				0.8125	0.338	221						0.6226	0.785	13	0.5989
2	28	1	53	0.6817	0.441	109					10	0.7703	0.735	27	0.5530
2	28	2	53	0.6835	0.440	110					10	0.7601	0.756	25	0.5410
2	28	3	53	0.6846	0.439	111					10	0.7428	0.788	23	0.5216
2				0.6833	0.440	110						0.7577	0.759	25	0.5385

The values in bold are values for the best selected model.
c = calibration set; F = Fischer ratio; n = number of compounds in the set; N_{act} = number of SMILES attributes involved in building up a model; probe = number of runs of the Monte Carlo method calculation; r = correlation coefficient; s = root-mean-standard error; t = training set; Thrsd = threshold; υ = test (validation) set.

(A < −0.5 < B < 0 < C < 0.5), electronegativity (A < 2.19 < B < 2.5 < C < 3 < D) and lipophilicity (A < −1 < B < −0.5 < C < −0.1 < D < 0.1 < E < 0.5 < F < 1 < G). The vertices of simplexes were marked by properties mentioned before.

After the differentiation step, all molecules were divided into fragments and all possible simplexes were calculated. Finally, the number of simplexes of definite type (for example, A-B-D-G) was used as a descriptor.

3. Results and discussion

Table 1 contains the data on the best statistical quality of the models obtained by using the CORAL approach with molecular GRAPHS and molecular SMILES using their extended connectivity. In the current study models based on EC0 in the HSG and S_k, SS_k in the NOSP, HALO, and PAIRS were selected as the best hybrid-based models. Statistical characteristics of the model for three splits of data obtained by the balance of correlations and by the classic scheme are reported in Table 1. These results were obtained with the threshold ranging from zero to three. How the number of epochs of the optimization influences the statistical quality of the model for the external test set was also studied. Fig. 1 shows the best model for pIC50 (Split 1, Probe 2, Threshold = 0).

This model is characterized below:

$$pIC_{50} = -0.587\ (\pm 0.081) + 0.114\ (\pm 0.001) * DCW\ (0, 30) \quad \text{Equation 4}$$

For a model with good external predictability, the cross-validation coefficient (R_m^2) value should be > 0.5. In the case of the model developed here the average R_m^2 value of the external set for all 11 compounds is about 0.70 and as such is quite satisfactory.

Equation 4 describes a satisfactory model, in view of two features: (1) the standard error for the external set is close to the training set, and (2) there are no influential outliers in either the training or the test sets, therefore all considered chemicals possess inhibitory activity.

Biological activity is related to the presence of molecular fragments with different roles: some increase, some reduce, and some do not have any effect on biological activity. These fragments can be distinguished by the optimization procedure. The approach under consideration requires the correlation coefficient between descriptors to be calculated with the correlation weight (CW) and inhibitory activity. Experimental and calculated values using Equation 4 values of pIC_{50} are displayed in Table 2. Table 3 contains the CW for calculation with Equation 4. SA_k is a symbol in SMILES notation. Subtraining (N_{train}), calibration

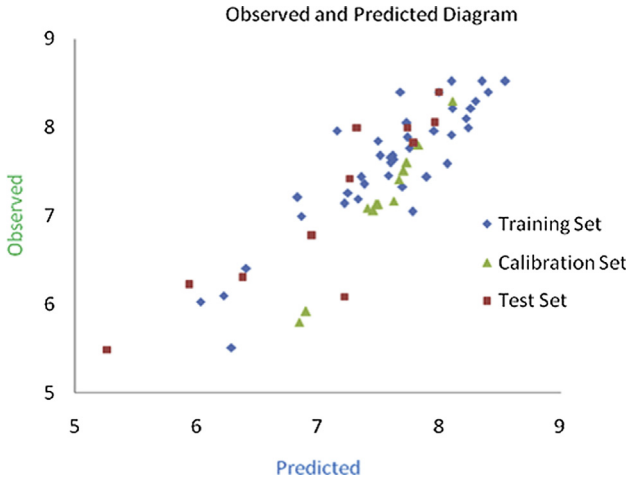

Fig. 1 – Graphical representation of the model calculated with Equation 4 (CORAL).

Table 2 – Experimental and calculated values of the activity pIC_{50} for 63 amino-substituted nitrogen heterocyclic ureas (CORAL).

Set	SMILES	DCW	Exp	Calc	Exp–Calc	ID
+	n1[nH]c2c(c1N)c(ccc2)c1ccc(cc1)NC(=O)Nc1ccccc1	69.700	7.190	7.331	−0.141	1
+	n1[nH]c2c(c1N)c(ccc2)c1ccc(cc1)NC(=O)Nc1cccc(c1)F	72.280	7.640	7.624	0.016	4
+	n1[nH]c2c(c1N)c(ccc2)c1ccc(cc1)NC(=O)Nc1ccc(cc1)C	76.495	7.920	8.103	−0.183	7
+	n1[nH]c2c(c1N)c(ccc2)c1ccc(cc1)NC(=O)Nc1cccc(c1)CC	77.901	8.220	8.263	−0.043	8
+	n1[nH]c2c(c1N)c(ccc2)c1ccc(cc1)NC(=O)Nc1cccc(c1)Cl	77.552	8.100	8.223	−0.123	9
+	n1[nH]c2c(c1N)c(ccc2)c1ccc(cc1)NC(=O)Nc1cc(ccc1)Br	69.940	7.440	7.358	0.082	10
+	n1[nH]c2c(c1N)c(ccc2)c1ccc(cc1)NC(=O)Nc1cccc(c1)C(F)(F)F	77.679	8.000	8.238	−0.238	11
+	n1[nH]c2c(c1N)c(ccc2)c1ccc(cc1)NC(=O)Nc1cccc(c1)O	68.989	7.260	7.250	0.010	12
+	n1[nH]c2c(c1N)c(ccc2)c1ccc(cc1)NC(=O)Nc1cc(ccc1F)C	72.783	8.400	7.681	0.719	13
+	n1[nH]c2c(c1N)c(ccc2)c1ccc(cc1)NC(=O)Nc1ccc(c(c1)C)F	75.566	8.400	7.998	0.402	14
+	n1[nH]c2c(c1N)c(ccc2)c1ccc(cc1)NC(=O)Nc1ccc(c(c1)F)C	74.699	7.440	7.899	−0.459	15
+	n1[nH]c2c(c1N)c(ccc2)c1ccc(cc1)NC(=O)Nc1cc(ccc1F)C(F)(F)F	73.656	7.050	7.781	−0.731	16
+	n1n(c2c(c1N)c(ccc2)c1ccc(cc1)NC(=O)Nc1cccc(c1)C)C	68.176	7.960	7.158	0.802	17
+	n1n(c2c(c1N)c(ccc2)c1ccc(cc1)NC(=O)Nc1cc(ccc1F)C)CCOC	60.506	5.510	6.287	−0.777	19
+	n1[nH]c2c(c1N)c(ccc2C)c1ccc(cc1)NC(=O)Nc1cc(ccc1)C	80.405	8.520	8.547	−0.027	20
+	n1[nH]c2c(c1N)c(ccc2OC)c1ccc(cc1)NC(=O)Nc1cccc(c1)C	76.195	7.590	8.069	−0.479	21
+	n1[nH]c2c(c1N)c(ccc2F)c1ccc(cc1)NC(=O)Nc1cccc(c1)C	78.237	8.300	8.301	−0.001	22
+	n1[nH]c2c(c1N)c(ccc2OCCN(CC)CC)c1ccc(cc1)NC(=O)Nc1cccc(c1)C	71.920	7.460	7.583	−0.123	25
+	n1[nH]c2c(c1N)c(ccc2OCCN1CCCC1=O)c1ccc(cc1)NC(=O)Nc1cc(ccc1)C	72.061	7.600	7.599	0.001	27
+	n1[nH]c2c(c1N)c(ccc2OCCOC)c1ccc(cc1)NC(=O)Nc1cccc(c1)C	72.237	7.680	7.619	0.061	28
+	n1[nH]c2c(c1N)c(ccc2CNN1CCOCC1)c1ccc(cc1)NC(=O)Nc1cc(ccc1)C	61.583	6.410	6.409	0.001	29
+	n1[nH]c2c(c1N)c(ccc2OCCOC)c1ccc(cc1)NC(=O)Nc1cc(ccc1)Cl	73.293	7.890	7.739	0.151	31
+	n1[nH]c2c(c1N)c(ccc2OCCOC)c1ccc(cc1)NC(=O)Nc1c(ccc(c1)C)F	71.308	7.680	7.514	0.166	32
+	n1[nH]c2c(c1N)c(ccc2OCCN1CCOCC1)c1ccc(cc1)NC(=O)Nc1c(ccc(c1)C)F	65.316	7.210	6.833	0.377	34
+	n1n(c2c(c1NC(=O)C)c(ccc2)c1ccc(cc1)NC(=O)Nc1cc(ccc1)C)C	48.445	4.920	4.917	0.003	35
+	c1c(c2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1cc(ccc1)C)N)Br	68.721	7.140	7.220	−0.080	37
+	c1cc2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1cc(ccc1)C)N	72.939	7.330	7.699	−0.369	38
+	c1c(c2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1cc(ccc1)C)N)c1ccccc1	72.105	7.660	7.604	0.056	39
+	c1c(c2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1cc(ccc1)C)N)c1ccsc1	75.226	7.960	7.959	0.001	42
+	c1c(c2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1cc(ccc1)C)N)c1ccc2c(c1)OCO2	73.421	7.770	7.754	0.016	44
+	c1c(c2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1cc(ccc1)C)N)c1nn(cc1)C	79.131	8.400	8.403	−0.003	46
+	c1c(c2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1cc(ccc1)C)N)C#N	58.251	6.030	6.031	−0.001	47
+	c1c(c2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1cc(ccc1)C)N)C(=O)NC	65.586	7.000	6.864	0.136	49
+	c1c(c2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1cc(ccc1)C)N)C(=O)N1CCN(CC1)C	59.955	6.100	6.224	−0.124	50
+	c1cc2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1cc(ccc1)C(F)(F)F)N	73.255	8.050	7.735	0.315	54
+	c1cc2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1cc(ccc1)Cl)N	70.138	7.360	7.381	−0.021	55
+	c1cc2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1c(ccc(c1)C(F)(F)F)F)N	71.148	7.850	7.496	0.354	57
+	c1c(c2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1cc(ccc1)C(F)(F)F)N)c1cn(nc1)C	78.680	8.520	8.351	0.169	58
+	c1c(c2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1cc(c(cc1)F)C(F)(F)F)N)c1cn(nc1)C~	76.573	8.220	8.112	0.108	61
–	n1[nH]c2c(c1N)c(ccc2)c1ccc(cc1)NC(=O)Nc1cccc(c1)C	76.495	8.520	8.103	0.417	2
–	n1[nH]c2c(c1N)c(ccc2)c1ccc(cc1)NC(=O)Nc1ccccc1F	70.364	7.090	7.407	−0.317	3
–	n1[nH]c2c(c1N)c(ccc2)c1ccc(cc1)NC(=O)Nc1ccc(cc1)F	72.280	7.170	7.624	−0.454	5
–	n1[nH]c2c(c1N)c(ccc2)c1ccc(cc1)NC(=O)Nc1ccccc1C	70.813	7.060	7.458	−0.398	6
–	n1[nH]c2c(c1N)c(ccc2OCCN1CCCC1)c1ccc(cc1)NC(=O)Nc1cccc(c1)C	73.014	7.510	7.708	−0.198	26
–	n1[nH]c2c(c1N)c(ccc2CN1CCN(CC1)C)c1ccc(cc1)NC(=O)Nc1cc(ccc1)C	65.923	5.920	6.902	−0.982	30
–	n1[nH]c2c(c1N)c(ccc2OCCN(CC)CC)c1ccc(cc1)NC(=O)Nc1c(ccc(c1)C)F	70.991	7.130	7.478	−0.348	33
–	c1cc2n(n1)c(c(c(n2)C)c1ccc(cc1)NC(=O)Nc1cc(ccc1)C)N	72.643	7.410	7.666	−0.256	40
–	c1c(c2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1cc(ccc1)C)N)C1CC1	74.085	7.800	7.829	−0.029	43
–	c1c(c2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1cc(ccc1)C)N)C(=O)OCC	73.238	7.600	7.733	−0.133	48
–	c1c(c2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1ccccc1)N)Br	65.435	5.800	6.847	−1.047	53
–	c1cc2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1c(ccc(c1)C)F)N	71.142	7.130	7.495	−0.365	56
–	c1c(c2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1c(ccc(c1)C)F)N)c1cn(nc1)C	76.567	8.300	8.111	0.189	60
#	n1n(c2c(c1N)c(ccc2)c1ccc(cc1)NC(=O)Nc1cccc(c1)C)CCO	57.426	6.220	5.937	0.283	18
#	n1[nH]c2c(c1N)c(ccc2Br)c1ccc(cc1)NC(=O)Nc1cccc(c1)C	69.665	8.000	7.327	0.673	23
#	n1[nH]c2c(c1N)c(ccc2OCCN(C)C)c1ccc(cc1)NC(=O)Nc1cccc(c1)C	69.109	7.420	7.264	0.156	24
#	n1n(c2c(c1N(C)C)c(ccc2)c1ccc(cc1)NC(=O)Nc1cc(ccc1)C)C	51.535	5.490	5.268	0.222	36
#	c1c(c2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1cc(ccc1)C)N)c1cccs1	73.770	7.820	7.794	0.026	41
#	c1c(c2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1cc(ccc1)C)N)c1cnccc1	75.269	8.050	7.964	0.086	45
#	c1c(c2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1cc(ccc1)C)N)C(=O)NCCN(CC)CC	61.312	6.300	6.378	−0.078	51
#	c1c(c2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1ccc(cc1)C)N)Br	68.721	6.090	7.220	−1.130	52
#	c1c(c2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1cc(ccc1)Cl)N)c1cn(nc1)C	75.563	8.400	7.997	0.403	59
#	c1c(c2n(n1)c(c(cn2)c1ccc(cc1)NC(=O)Nc1cc(ccc1)F)N)c1cn(nc1)C	73.281	8.000	7.738	0.262	62
#	c12c(c(c(cc1)c1ccc(cc1)NC(=O)Nc1cccc(c1)C)N)nccn2	66.390	6.770	6.955	−0.185	63

CORAL = CORrelations And Logic; SMILES = simplified molecular input-line entry systems; DCW = descriptor of correlation weights.
Exp and Calc are experimental and calculated pIC50; "+", "–", and "#" are indicators for the training, calibration, and test sets, respectively.

Table 3 – Correlation weights for calculation of DCW (SMILES) used in Eq. (1).

SA_k	CW (SA_k)	N_{train}[a]	N_{calib}[a]	N_{test}[a]
#.	−2.44150	1	0	0
(...(..	7.79887	6	0	0
(.	−0.95894	39	13	11
++++B2–B3==	−1.52444	1	0	0
++++F—B2==	0.26281	14	5	1
++++F—N===	−0.16006	14	5	1
++++F—O===	0.20794	14	5	1
++++CL–N===	0.33213	3	0	1
++++CL–O===	1.13663	3	0	1
++++Br–B2==	−0.13863	2	1	2
++++Br–N===	−2.68450	2	1	2
++++Br–O===	0.03606	2	1	2
++++Cl–B2==	2.87200	3	0	1
++++N—B2==	8.05269	39	13	11
++++N—B3==	−3.19050	1	0	0
++++N—O===	9.31931	39	13	11
++++N—S===	−0.19131	1	0	1
++++O—B2==	8.84956	39	13	11
++++O—B3==	−4.31250	1	0	0
++++O—S===	0.99619	1	0	1
++++S—B2==	−2.30469	1	0	1
1...(..	3.79988	39	13	11
1.	−0.79406	39	13	11
2...(..	1.41125	29	10	8
2.	−1.97956	39	13	11
2...1..	0.99500	0	0	1
=...(..	1.90125	39	13	11
=.	0.57331	39	13	11
=...1..	5.67387	1	0	0
C...#..	−3.43750	1	0	0
C...(..	0.29106	39	13	11
C.	−0.42569	39	13	11
C...1..	1.26863	4	4	0
C...2..	5.18650	2	1	0
C...C..	1.87500	10	5	3
BOND10000000	9.25481	38	13	11
BOND11000000	−2.22175	1	0	0
F...(..	−0.57613	14	4	1
F.	−0.28706	14	5	1
F...1..	4.16887	3	1	0
F...2..	7.23338	1	0	0
EC0-C...1...	2.28525	29	10	11
EC0-C...2...	−0.04388	39	13	11
EC0-C...3...	−2.05969	39	13	11
EC0-C...4...	7.18350	6	0	0
EC0-F...1...	−1.51262	14	5	1
EC0-Br..1...	−0.00681	2	1	2
EC0-Cl..1...	1.99719	3	0	1
EC0-N...1...	15.74219	38	13	10
EC0-N...2...	−1.69731	39	13	11
EC0-N...3...	−3.86419	21	9	9
EC0-O...1...	−5.31931	39	13	11
EC0-O...2...	1.61137	10	3	1
EC0-s...2...	1.09856	1	0	1
H.	1.07331	22	7	2
Br..(..	−0.09175	2	1	2
Br	−1.51944	2	1	2
Br..2..	1.00200	0	0	1
Cl..(..	−3.56550	3	0	1
Cl	0.43450	3	0	1
N...#..	−2.11037	1	0	0
N...(..	−0.65725	39	13	11
N.	−1.94531	39	13	11
N...1..	−2.28325	26	7	4
N...C..	−2.36919	39	13	11
N...N..	−3.82512	1	0	0
O...(..	2.47275	39	13	11
O.	−2.50881	39	13	11
O...2..	3.87100	8	2	1
O...=..	3.97175	39	13	11
O...C..	−1.99800	10	3	2
NOSP11000000	9.06450	39	13	11
[.	1.98438	22	7	2
[...1..	3.74019	22	7	2
[...H..	2.99800	22	7	2
c...(..	1.69150	39	13	11
c.	0.98838	39	13	11
c...1..	0.26863	39	13	11
c...2..	0.58394	39	13	11
c...N..	9.94150	39	13	11
c...[..	4.05669	22	7	2
c...c..	−1.39544	39	13	11
n...(..	1.18850	17	6	9
n.	−0.35256	39	13	11
n...1..	6.86037	39	13	10
n...2..	3.49619	14	6	7
n...H..	0.54488	22	7	2
n...[..	1.73238	22	7	2
n...c..	1.68369	14	5	7
n...n..	−1.56450	1	0	0
s.	−0.36238	1	0	1
s...1..	1.00200	0	0	1
s...c..	0.79387	1	0	1

CW = correlation weight; DCW = descriptor of correlation weights; SMILES = simplified molecular input-line entry systems.
[a] The N_{train}, N_{calib}, and N_{test} are the frequencies of SA_k in the training, calibration, and test sets, respectively.

(N_{calib}) and test sets (N_{test}) represent distribution of structural attributes.

The results obtained by SiRMS are summarized in Table 4. In model 1S the fragments representing tetratomic bonded simplexes were used. In model 2S tetratomic unbound simplexes were used. In model 3S unbound fragments of the size 2–5 were used. Each model consists of nine descriptors. As seen in Table 4, all models have similar statistical characteristics. Despite this, it is necessary to consider the second model for further interpretation since the first and the third models do not distinguish structural isomers. Thus, nine significant descriptors were combined into four groups: type of atom, lipophilicity, van-der-Waals interactions, and partial charges. The relative influences (%) are presented in Fig. 2.

Three descriptors of atom type reflect differences among functional groups located in the same place of the molecule. The descriptor of partial charges describes differences for aromatic substitution. Lipophilicity reflects the impact of nonaromatic connectors between aromatic parts of molecules. A set of van-der-Waals-related descriptors includes four descriptors. They describe the influence of aromatic substitution, and the impact of functional groups. A plot of experimentally determined versus predicted log values is presented in Fig. 3.

Table 4 – Summarized statistical evaluation of each model developed by the SiRMS approach.

Model (split)	$R^2_{training}$	$s_{training}$	q^2	$s_{cross-validation}$	R^2_{test}	s_{test}
1S	0.86	0.31	0.81	0.37	0.75	0.47
2S	0.84	0.33	0.79	0.39	0.70	0.50
3S	0.82	0.35	0.76	0.42	0.72	0.49

q = LOO cross-validation coefficient; R^2 = correlation coefficient; s = standard error; SiRMS = simplex representation of molecular structure.

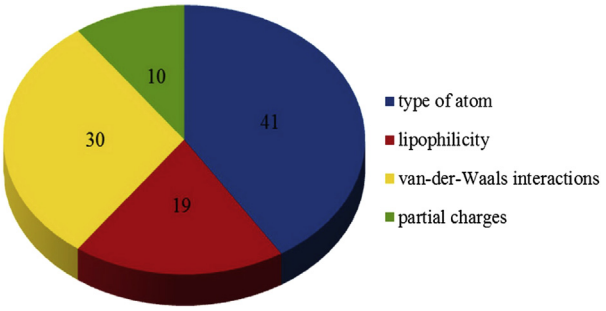

Fig. 2 – Diagram of relative influence (%) of various groups of SiRMS descriptors.

It can be noted that both approaches applied in this study (SMILES-based optimal descriptors and SiRMS) deliver good performance in prediction of KDR inhibitory activity by amino-substituted heterocyclic urea derivatives. As seen in Table 2 and Table 4, both approaches display similar results on average.

4. Conclusion

A structure–activity relationship analysis was performed for a set of amino-substituted nitrogen heterocyclic urea derivatives. Two approaches were applied: the SMILES-based optimal descriptors approach (CORAL) and the fragment-based SiRMS approach. In the case of the SMILES-based optimal descriptors approach, three various splits of the experimental data into subtraining set, calibration set, and test set were examined. Comparison of the classic scheme of building up the model and balance of correlation (BC) scheme show that the balance scheme is characterized by more robust predictions than the classic scheme for the pIC_{50} of the studied compounds. The SiRMS approach was examined for three various splits of the descriptors set. Comparison of the SMILES-based optimal descriptors and SiRMS approaches has confirmed a good performance of both approaches in prediction of KDR inhibitory activity (pIC_{50}) of amino-substituted nitrogen heterocyclic urea derivatives. Both methods are quite fast and reliable and possess comparable statistical quality.

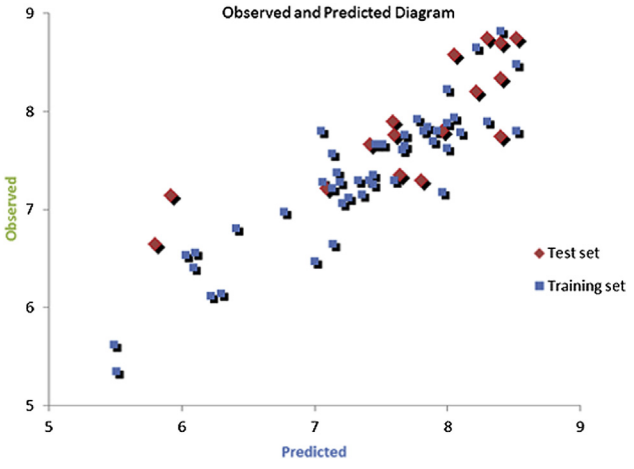

Fig. 3 – Plot of experimental (observed) versus predicted log values, SiRMS approach.

Conflicts of interest

The authors have no financial or commercial conflicts of interest.

Acknowledgments

This work was financially supported by Erciyes University BAP of Turkey (Grant No; FBA-12-3578) and by NSF CREST Interdisciplinary Nanotoxicity Center NSF-CREST – Grant #HRD-0833178; AAT thanks EU project PROSIL funded under the LIFE program (project LIFE12ENV/IT/000154). The authors also thank the Mississippi Center for Supercomputer Research (Oxford, MS) for a generous allotment of computer time.

REFERENCES

[1] Boyer SJ. Small molecule inhibitors of KDR (VEGFR-2) kinase: an overview of structure activity relationships. Curr Top Med Chem 2002;2:973–1000.
[2] Ferrara N, Hillan KJ, Gerber HP, Novotny W. Discovery and development of bevacizumab, an anti-VEGF antibody for treating cancer. Nat Rev Drug Discov 2004;3:391–400.
[3] Sosnowska A, Barycki M, Jagiello K, Haranczyk M, Gajewicz A, Kawai T, Suzuki N, Puzyn T. Predicting enthalpy of vaporization for persistent organic pollutants with quantitative structure-property relationship (QSPR) incorporating the influence of temperature on volatility. Atmos Environ 2014;87:10–8.
[4] Ahmed L, Rasulev B, Turabekova M, Leszczynska D, Leszczynski J. Receptor- and ligand-based study of fullerene analogues: comprehensive computational approach including quantum-chemical, QSAR and molecular docking simulations. Org Biomol Chem 2013;11:5798–808.
[5] Toropov AA, Toropova AP, Rasulev BF, Benfenati E, Gini G, Leszczynska D, Leszczynski J. CORAL: QSPR modeling of rate constants of reactions between organic aromatic pollutants and hydroxyl radical. J Comp Chem 2012;33:1902–6.

[6] Toropova AP, Toropov AA, Rasulev BF, Benfenati E, Gini G, Leszczynska D, Leszczynski J. QSAR models for ACE-inhibitor activity of tri-peptides based on representation of the molecular structure by graph of atomic orbitals and SMILES. Struct Chem 2012;23:1873–8.
[7] Petrova T, Rasulev BF, Toropov AA, Leszczynska D, Leszczynski J. Improved model for fullerene C_{60} solubility in organic solvents based on quantum-chemical and topological descriptors. J Nanopart Res 2011;13:3235–47.
[8] Rasulev B, Turabekova M, Gorska M, Kulig K, Bielejewska A, Lipkowski J, Leszczynski J. Use of quantitative structure-enantioselective retention relationship for the liquid chromatography chiral separation prediction of the series of pyrrolidin-2-one compounds. Chirality 2012;24:72–7.
[9] Toropov AA, Toropova AP, Puzyn T, Benfenati E, Gini G, Leszczynska D, Leszczynski J. QSAR as a random event: modeling of nanoparticles uptake in PaCa2 cancer cells. Chemosphere 2013;92:31–7.
[10] Turabekova MA, Vinogradova VI, Werbovetz KA, Capers J, Rasulev BF, Levkovich MG, Rakhimov SB, Abdullaev ND. Structure-activity relationship investigations of leishmanicidal N-benzylcytisine derivatives. Chem Biol Drug Des 2011;78:183–9.
[11] Rasulev BF, Kušic H, Leszczynska D, Leszczynski J, Koprivanac N. QSAR modeling of acute toxicity on mammals for aromatic compounds: The case study using oral LD_{50} for rats. J Environ Monit 2010;12:1037–44.
[12] Puzyn T, Gajewicz A, Rybacka A, Haranczyk M. Global versus local QSPR models for persistent organic pollutants: balancing between predictivity and economy. Struct Chem 2011;22:873–84.
[13] Turabekova MA, Rasulev BF, Dzhakhangirov FN, Leszczynska D, Leszczynski J. Aconitum and delphinium alkaloids of curare-like activity. QSAR analysis and molecular docking of alkaloids into AChBP. Eur. J Med Chem 2010;45:3885–94.
[14] Kušić H, Rasulev B, Leszczynska D, Leszczynski J, Koprivanac N. Prediction of rate constants for radical degradation of aromatic pollutants in water matrix: a QSAR study. Chemosphere 2009;75:1128–34.
[15] Paukku Y, Rasulev BF, Syrov V, Khushbaktova Z, Leszczynski J. Structure-hepatoprotective activity relationship study of sesquiterpene lactones: A QSAR analysis. Int J Quant Chem 2009;109:17–27.
[16] Rasulev BF, Saidkhodzhaev AI, Nazrullaev SS, Akhmedkhodzhaeva KS, Khushbaktova ZA, Leszczynski J. Molecular modeling and QSAR analysis of the estrogenic activity of terpenoids isolated from Ferula plants. SAR QSAR Environ Res 2007;18:663–73.
[17] Basak SC, Gute BD, Lucic B, Nikolic S, Trinajstic N. A comparative QSAR study of benzamidines complement–inhibitory activity and benzene derivatives acute toxicity. Comp Chem 2000;24:181–91.
[18] Artemenko AG, Muratov EN, Kuz'min VE, Muratov NN, Varlamova EV, Kuz'mina AV, Gorb LG, Golius A, Hill FC, Leszczynski J, Tropsha A. QSAR analysis of the toxicity of nitroaromatics in Tetrahymena pyriformis: structural factors and possible modes of action. SAR QSAR Environ Res 2011;22:575–601.
[19] Lu X, Chen Y, You Q. Pharmacophore guided 3D-QSAR CoMFA analysis of amino substituted nitrogen heterocycle ureas as KDR inhibitors. QSAR Comb Sci 2009;28:1524–36.
[20] Toropov AA, Benfenati E. SMILES in QSPR/QSAR modeling: Results and perspectives. Curr Drug Discov Technol 2007;4:77–116.
[21] Toropov AA, Benfenati E. Additive SMILES-based optimal descriptors in QSAR modelling bee toxicity: Using rare SMILES attributes to define the applicability domain. Bioorg Med Chem 2008;26:4801–9.
[22] Toropov AA, Toropova AP, Benfenati E. Simplified molecular input line entry system-based optimal descriptors: quantitative structure–activity relationship modeling mutagenicity of nitrated polycyclic aromatic hydrocarbons. Chem Biol Drug Des 2009;73:515–25.
[23] Toropov AA, Toropova AP, Martyanov SE, Benfenati E, Gini G, Leszczynska D, Leszczynski J. CORAL: Predictions of rate constants of hydroxyl radical reaction using representation of the molecular structure obtained by combination of SMILES and Graph approaches. Chemometr Intell Lab Syst 2012;112:65–70.
[24] Toropov AA, Toropova AP. Prediction of heteroaromatic amine mutagenicity by means of correlation weighting of atomic orbital graphs of local invariants. J Mol Struct (Theochem) 2001;538:287–93.
[25] CORAL software (CORrelations And Logic), http://www.insilico.eu/coral [accessed 02.04.15].
[26] Ognichenko LN, Kuz'min VE, Gorb L, Hill FC, Artemenko AG, Polischuk PG, Leszczynski J. QSPR prediction of lipophilicity for organic compounds using random forest technique on the basis of simplex representation of molecular structure. Mol Inform 2012;31:273–80.
[27] Kuz'min VE, Polishchuk PG, Artemenko AG, Andronati SA. Interpretation of QSAR models based on random forest methods. Mol Inform 2011;30:593–603.
[28] Kovdienko NA, Polishchuk PG, Muratov EN, Artemenko AG, Kuz'min VE, Gorb L, Hill F, Leszczynski J. Application of random forest and multiple linear regression techniques to QSPR prediction of an aqueous solubility for military compounds. Mol Inform 2010;29:394–406.
[29] Ognichenko LN, Kuz'min VE, Artemenko AG. New structural descriptors of molecules on the basis of symbiosis of the informational field model and simplex representation of molecular structure. QSAR Comb Sci 2009;28:939–45.