European Journal of Medicinal Chemistry 44 (2009) 2544–2551

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Original article

QSPR modeling bioconcentration factor (BCF) by balance of correlations

A.A. Toropov [a,b,*], A.P. Toropova [a,b], E. Benfenati [b]

[a] Institute of Geology and Geophysics, Khodzhibaev Street 49, 100041 Tashkent, Uzbekistan
[b] Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy

ARTICLE INFO

Article history:
Received 14 July 2008
Received in revised form
18 January 2009
Accepted 22 January 2009
Available online 31 January 2009

Keywords:
QSPR
SMILES
Bioconcentration factor
Optimal descriptor
Balance of correlations

ABSTRACT

In many cases, quantitative structure–property/activity relationships (QSPRs/QSARs) are built according to the following scheme: (1) a split of the chemicals into a training and test sets; (2) selection of a model satisfactory for the training set; (3) validation of the model with the test set. Balance of correlations is an approach we used with the following scheme: (1) a split into a subtraining set, a calibration set, and a test set; (2) selection of a model that is satisfactory for both the subtraining and calibration sets; (3) validation of the model with the external test set. Comparison of these schemes for optimal descriptors calculated with simplified molecular input line entry system (SMILES) has shown that the use of the correlation balance gives better prediction of bioconcentration factor for the test set. The calculations were carried out for three random splits into a subtraining, a calibration, and a test set.

1. Introduction

The bioconcentration factor (BCF) is useful to characterize the environmental behaviour of a chemical substance and in particular if a chemical is likely to accumulate. BCF defines the ratio between the concentration in the organism and the medium. Typically, BCF is measured in fish, but other organisms can be used. Different thresholds are defined for BCF. For instance, a chemical may be considered potentially bioaccumulative if the BCF is greater than 500 [1]. However, the EU legislation REACH defines a threshold of 2000 for a bioaccumulative compound, and of 5000 for a very bioaccumulative [2] compound. The same legislation solicits to assess the possibility of using alternative methods to evaluate chemical properties, in order to reduce the animal use. Quantitative structure–property/activity relationships (QSPRs/QSARs) are also mentioned in this legislation, and QSAR methods for BCF have been developed and used for regulatory purposes in the USA [3]. To validate the developed model is important. Indeed the model's utility depends on the possibility to quantify the model's capability to predict unknown chemicals with a clear degree of certainty [4]. The absence of this possibility leads to serious criticism of QSPR/QSAR results [5].

The genesis of the QSAR and components such as the descriptors used for the model has both heuristic and technical significance. For instance, descriptors such as LOMO and HUMO energies [6] are available for a limited number of laboratories. On the other hand, the number of internet databases providing simplified molecular input line entry system (SMILES) [7–10] is gradually increasing. Thus, the search for SMILES-based methods for QSPR/QSAR analyses becomes useful for both theoretical and practical aspects.

A good model for the training set can be a poor one for an external test set. A method of correlation balance described in Ref. [11] was effective to avoid this situation.

The aim of the present study is an estimation of predictive ability of the models of **log BCF** which have been obtained by means of the SMILES-based optimal descriptors calculated by the method of correlation balance [11]. The significant number of articles which is dedicated to the bioconcentration factor indicates importance of this endpoint [12–26].

2. Method

2.1. Data

In this study the numerical data on the bioconcentration factor (**log BCF**) was taken from Ref. [27].

An usual practice for the QSAR is scheme: (1) a split of an available experimental data into two sets: a training set and a test set; (2) construction of a model using the training set; and (3) an evaluation of the predictive potential of the model, generated with the training set, for the test set.

* Corresponding author. Institute of Geology and Geophysics, Khodzhibaev Street 49, 100041 Tashkent, Uzbekistan.
E-mail address: aatoropov@yahoo.com (A.A. Toropov).

0223-5234/$ – see front matter © 2009 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2009.01.023

Table 1
List of the global SMILES attributes.

SA_k	Definition	Reason
!-01_____ !-02_____ … !000_____ !001_____ … !010_____	Difference between the numbers of 'c' (lowercase letter) and the number of 'C' (capital letter)	Carbons in sp^2 state generate a rigid molecular fragment. Carbons in sp^3 state generate a flexible molecular fragment. The chlorine atom also is a provider of the 'C' (capital letter), but chlorine also is a flexible (not rigid) fragment. Thus this difference is an indicator of behaviour of a molecule in mechanical aspect (presence/absence of rigid and flexible fragments)
(000_____ (001_____ … (025_____	Number of branches in the molecular skeleton. (000_____ is indicator of absence of the branching	Branching has influence on many physico-chemical properties (normal boiling and melting points, viscosity, density, etc.)
1000_____ 1002_____ … 1006_____ 2000_____ 2002_____ … 5002_____	Numbers of cycles encoded in the SMILES with '1', '2', … '5' The 1000_____, 2000_____, …5000____ are indicators of the absence of the cycles	Cycles have influence on many physico-chemical and biological parameters
=000_____ =001_____ … =006_____	Number of the double covalent bonds in molecule. = 000____ is an indicator of absence of double bonds; = 001 is an indicator of presence of one double bond, etc.	Double bonds have a effect upon many biochemical phenomena
C000_____ C001_____ … C027_____	Number of carbon atoms in the sp^3 electronic state	Each chemical element theoretically can have an effect to any endpoint
F000_____ … F027_____	Number of fluorine atoms in molecule	Each chemical element theoretically can have an effect to any endpoint
Br00_____ … Br10_____	Number of bromine atoms in molecule	Each chemical element theoretically can have an effect to any endpoint
Cl00_____ … Cl08_____	Number of chlorine atoms in molecule	Each chemical element theoretically can have an effect to any endpoint
N000_____ … N006_____	Number of nitrogen atoms in molecule	Each chemical element theoretically can have an effect to any endpoint
O000_____ … O010_____	Number of oxygen atoms in molecule	Each chemical element theoretically can have an effect to any endpoint
P000_____ P001_____	Number of phosphorus atoms in molecule	Each chemical element theoretically can have an effect to any endpoint
S000_____ … S003_____	Number of sulphur atoms in molecule	Each chemical element theoretically can have an effect to any endpoint
c000_____ … c022_____	Number of carbon atoms in the sp^2 electronic state	Each chemical element theoretically can have an effect to any endpoint

As an alternative, the training set can be split into two sets: a subtraining set and a calibration set and the model obtained with the subtraining set can be preliminarily evaluated with substances of the calibration set. In fact, the classical training phase gives a maximum of the correlation coefficient between the actual and predicted values for an endpoint of interest. Using the subtraining and calibration sets, one can calculate the combinatorial measure of correlation between actual and predicted values, as the following:

$$B = R_s^2 + R_c^2 - \mathrm{abs}\left(R_s^2 - R_c^2\right) \quad (1)$$

where R_s and R_c are correlation coefficients between the actual and predicted values of an endpoint for the subtraining set and calibration set, respectively. It is likely that a model, obtained by the scheme [subtraining → calibration → prediction] (i.e., by the correlation balance), is more robust in comparison with the model, obtained by the 'classic' scheme [training → test], without the preliminary calibration [11].

One-variable models examined in this study are based on optimal descriptors, calculated as the following:

$$\mathrm{DCW(LimN)} = \prod \mathrm{CW}(SA_k) \quad (2)$$

where $\mathbf{SA_k}$ is an attribute of the SMILES notation; the $\mathbf{CW}(\mathbf{SA_k})$ is the correlation weight for the $\mathbf{SA_k}$; the **LimN** described in Ref. [11] is a parameter that defines two categories: "rare SMILES attribute" and "non rare SMILES attribute". In other words, the **LimN** is the minimal number of $\mathbf{SA_k}$ in the training (or subtraining for case of the correlation balance) set. For instance, if **LimS** = 4 and only three (or less) $\mathbf{SA_k}$ exist in the training (or subtraining) set then this attribute should be classified as rare. The correlation weights of the rare SMILES attributes are equal to 1. The correlation weights of the non-rare SMILES attributes are calculated by optimization of the Monte Carlo method: the target function can be the correlation coefficient between **DCW**(**limN**) and **log BCF** for the training set or the criterion that is calculated with Eq. (1) for the subtraining set and calibration set.

The use of different **LimNs** is accompanied by different statistical qualities of the models. If **LimN** tends to zero, the correlation coefficient between **DCW**(**limN**) and **log BCF** asymptotically tends to unit for the training set. An increase of the **LimN** leads to a decrease of the correlation coefficient for the training set, owing to the decrease of the number of optimized parameters. However, the increase of the **limN** can lead to an increase of the correlation coefficient for the test set [11,28].

The local and global SMILES attributes are used in this study. The local attributes are constructed with elements of the SMILES. The element can be one character of the SMILES or two characters, which cannot be separated, without the loss of the physical sense: for instance, 'Cl', 'Br', 'Si', etc. Combinations of one, two and three SMILES elements are used as local SMILES attributes. Combinations of two (AB) and three (ABC) SMILES elements are organized according to ASCII codes. In other words these local attributes have only one version in the list of SMILES attributes (e.g., AB, not both AB and BA; similarly ABC, not ABC and CBA: it should be noted that the positions of A and C can be changed, B in any case has middle position). The global SMILES attributes are calculated with the SMILES. These are the number of a chemical element (C, O, N, etc.) or other molecular features, such as branching in molecular skeleton, cycles, etc. Table 1 contains the list of global SMILES attributes.

The split into training and test sets has an influence on the results of the QSPR/QSAR modeling. Hence, the robustness of the model should be estimated for several splits.

Three splits into subtraining, calibration, and test set for modeling by the scheme [subtraining → calibration → test] are examined. Supplementary material section contains CAS numbers of substances randomly extracted in the external test sets into splits A, B, and C. For modeling by scheme [training → test], the association of the subtraining and calibration sets was used as the united training set.

Optimization of the correlation weights has been carried out by algorithm based on the Monte Carlo method [11]. For each intermediate list of the correlation weights, the **DCW**(**limN**) values for all compounds have been calculated. One can, using these data, calculate the R_s and R_c for the [subtraining → calibration → test] scheme or the correlation coefficient for the united subtraining set

Table 2
The statistical characteristics of the QSAR models (the averages for three probes of the Monte Carlo optimization).

LimS	N_{act}	Subtraining set, $n = 193$			Calibration set, $n = 175$			Test set, $n = 105$		
		r^2	s	F	r^2	s	F	r^2	s	F
Split A [subtraining→calibration→test] scheme										
10	179	0.8136	0.575	834	0.8364	0.573	885	0.6895	0.689	229
11	172	0.8053	0.588	790	0.8306	0.580	850	0.7055	0.668	247
12	164	0.7928	0.606	731	0.8308	0.581	850	0.6877	0.687	227
13	152	0.7762	0.630	663	0.8298	0.584	843	0.7346	0.630	285
14	148	0.7627	0.649	614	0.8205	0.601	791	0.7418	0.619	296
15	144	0.7637	0.647	618	0.8165	0.608	770	0.7397	0.620	293
16	142	0.7583	0.655	599	0.8162	0.607	769	0.7502	0.602	310
17	136	0.7585	0.654	600	0.8135	0.612	755	**0.7761**	0.575	358
18	134	0.7581	0.655	599	0.8069	0.622	723	0.7728	0.576	354
19	129	0.7514	0.664	577	0.8124	0.616	749	0.7690	0.581	345
20	109	0.7347	0.686	529	0.8017	0.633	700	0.7539	0.609	320
Split A [training→test] scheme										
10	250	0.8785	0.478	2648				0.5275	0.855	115
11	240	0.8747	0.485	2556				0.6076	0.759	160
12	228	0.8676	0.499	2399				0.5872	0.782	147
13	217	0.8557	0.521	2171				0.6323	0.739	178
14	209	0.8480	0.535	2043				0.6744	0.689	215
15	204	0.8508	0.530	2088				0.6351	0.735	181
16	197	0.8372	0.553	1882				**0.6834**	0.680	222
17	195	0.8372	0.553	1884				0.6766	0.686	216
18	195	0.8370	0.554	1882				0.6717	0.696	212
19	189	0.8310	0.564	1800				0.6666	0.710	206
20	182	0.8243	0.575	1718				0.6802	0.702	219
Split B [subtraining→calibration→test] scheme										
10	177	0.7859	0.637	701	0.8370	0.551	889	0.7411	0.625	295
11	171	0.7908	0.630	722	0.8277	0.566	832	0.7741	0.585	353
12	157	0.7762	0.652	663	0.8253	0.569	818	0.7919	0.552	392
13	153	0.7755	0.653	660	0.8259	0.567	821	0.7840	0.562	374
14	152	0.7729	0.656	650	0.8199	0.577	788	**0.7968**	0.543	404
15	146	0.7719	0.658	646	0.8206	0.578	791	0.7921	0.549	393
16	127	0.7673	0.664	630	0.8032	0.604	706	0.7532	0.603	315
17	123	0.7705	0.660	642	0.7965	0.615	677	0.7622	0.590	330
18	116	0.7694	0.661	637	0.7791	0.642	610	0.7475	0.612	306
19	113	0.7643	0.669	619	0.7806	0.638	616	0.7681	0.582	342
20	111	0.7607	0.674	607	0.7752	0.646	597	0.7705	0.578	346
Split B [training→test] scheme										
10	241	0.8620	0.508	2287				0.6186	0.799	167
11	233	0.8522	0.526	2110				0.6663	0.735	206
12	223	0.8506	0.529	2085				0.7091	0.681	253
13	215	0.8445	0.539	1989				0.6816	0.719	221
14	203	0.8343	0.557	1843				0.6839	0.718	226
15	201	0.8262	0.570	1740				0.6662	0.737	208
16	198	0.8253	0.572	1729				0.6348	0.784	181
17	193	0.8168	0.585	1632				0.7312	0.636	281
18	190	0.8133	0.591	1595				0.7130	0.664	256
19	188	0.8163	0.586	1628				**0.7271**	0.642	279
20	184	0.8105	0.595	1567				0.7087	0.668	251
Split C [subtraining→calibration→test] scheme										
10	183	0.7881	0.623	710	0.8480	0.554	966	0.7340	0.686	285
11	176	0.7869	0.625	705	0.8413	0.565	917	**0.7831**	0.619	372
12	169	0.7745	0.643	656	0.8416	0.565	920	0.7640	0.640	334
13	164	0.7736	0.644	653	0.8346	0.573	873	0.7636	0.642	333
14	160	0.7621	0.660	613	0.8324	0.584	859	0.7722	0.626	350
15	154	0.7670	0.654	629	0.8339	0.577	869	0.7636	0.641	333
16	147	0.7625	0.660	613	0.8302	0.583	846	0.7773	0.620	360
17	142	0.7523	0.674	580	0.8239	0.595	809	0.7778	0.617	361
18	131	0.7569	0.668	595	0.8189	0.605	783	0.7621	0.642	330
19	122	0.7454	0.683	559	0.8201	0.602	789	0.7603	0.641	327
20	115	0.7380	0.693	538	0.8188	0.600	782	0.7657	0.633	337
Split C [training→test] scheme										
10	246	0.8583	0.505	2216				0.7039	0.724	245
11	235	0.8567	0.508	2189				0.6982	0.738	239
12	227	0.8503	0.519	2079				0.7003	0.734	241
13	221	0.8441	0.529	1983				0.7064	0.742	248
14	212	0.8307	0.552	1796				0.7181	0.705	264
15	208	0.8336	0.547	1835				0.7080	0.721	250
16	200	0.8183	0.571	1649				0.7280	0.703	276
17	193	0.8180	0.572	1645				**0.7359**	0.685	287
18	188	0.8140	0.578	1603				0.7236	0.707	270

Table 2 (*continued*)

LimS	N_{act}	Subtraining set, $n = 193$			Calibration set, $n = 175$			Test set, $n = 105$		
		r^2	s	F	r^2	s	F	r^2	s	F
19	187	0.8154	0.576	1617				0.7214	0.711	267
20	179	0.8118	0.582	1579				0.7341	0.687	285

Best models are indicated by bold.

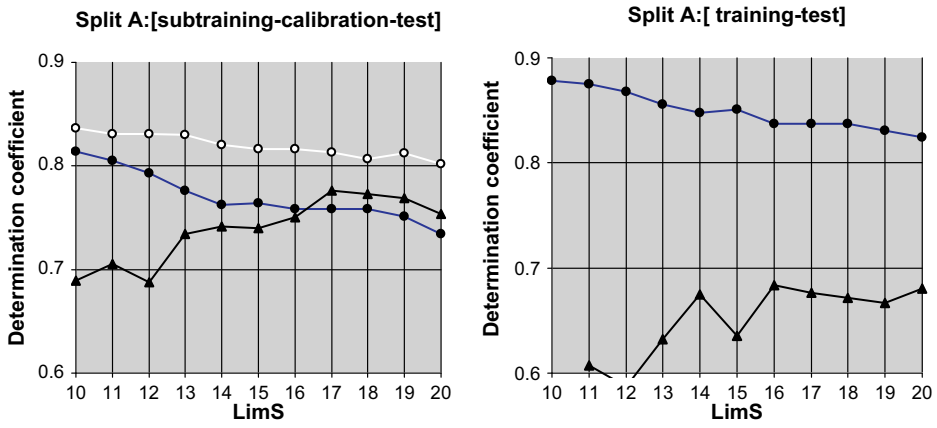


Fig. 1. The correlation coefficient versus the **LimN**, for training/subtraining (circles), calibration (white line), and test (triangles) sets, for the splits A.

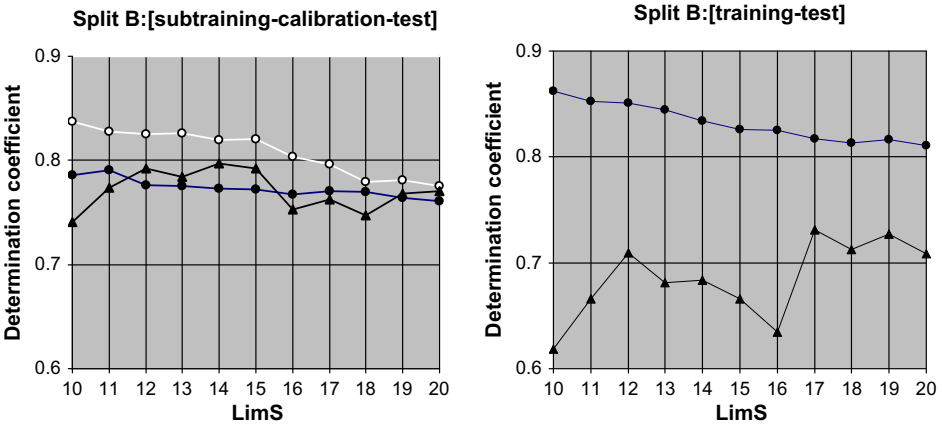


Fig. 2. The correlation coefficient versus the **LimN**, for training/subtraining (circles), calibration (white line), and test (triangles) sets, for the splits B.

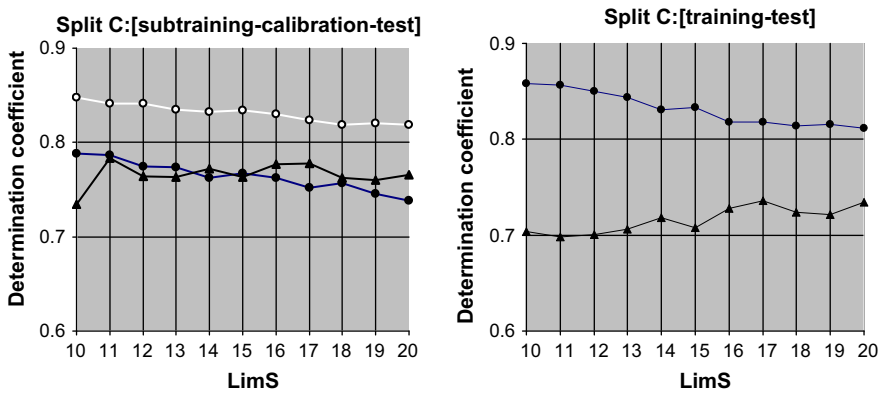


Fig. 3. The correlation coefficient versus the **LimN**, for training/subtraining (circles), calibration (white line), and test (triangles) sets, for the splits C.

Table 3
Correlation weights for the **DCW**(**14**) calculation.

SA_k	$CW(SA_k)$	ID	No.	NS_{TRN}	NS_{CLB}	NS_{TST}
!004_____	0.9990760	26	1	22	15	9
!005_____	1.0011436	27	2	15	12	11
(000_____	0.9925968	44	3	27	16	13
(001_____	0.9975935	45	4	34	43	30
(002_____	0.9998299	46	5	55	39	25
(003_____	1.0012398	47	6	26	32	14
(004_____	1.0045484	48	7	23	22	8
(_____	0.9955638	61	8	1004	964	514
+_____	1.0023668	62	9	15	19	11
1000_____	1.0063275	63	10	44	33	19
1002_____	1.0099810	64	11	102	93	64
1004_____	1.0051069	65	12	42	46	19
2000_____	1.0055077	67	13	155	136	81
2002_____	1.0050477	68	14	30	36	20
−_____	1.0022517	70	15	15	15	12
3000_____	1.0133248	71	16	184	163	98
4000_____	1.0014690	74	17	190	173	104
5000_____	0.9961318	76	18	193	174	105
1_____	1.0014800	78	19	402	388	222
2_____	0.9985624	79	20	92	84	56
3_____	1.0041011	80	21	18	26	14
=000_____	1.0061254	83	22	120	115	60
=001_____	1.0084525	84	23	35	33	28
=002_____	1.0055796	85	24	24	18	10
=_____	1.0058437	90	25	136	100	73
C000_____	0.9978814	91	26	27	31	15
C001_____	0.9931549	92	27	17	17	16
C002_____	0.9975650	93	28	35	28	18
C004_____	0.9991559	95	29	22	25	8
C005_____	0.9969955	96	30	15	13	8
F000_____	1.0020780	113	31	189	171	103
C_____	1.0013833	121	32	775	558	387
Br00_____	0.9952189	122	33	182	167	99
Cl00_____	1.0151967	130	34	130	112	71
Cl01_____	1.0058571	131	35	16	12	9
Cl04_____	0.9987757	134	36	15	13	4
Br_____	1.0001185	139	37	27	33	14
Cl_____	1.0061077	140	38	190	191	91
F_____	1.0048141	141	39	27	60	5
N000_____	1.0129657	142	40	132	121	75
N001_____	1.0033796	143	41	45	26	18
O000_____	1.0088745	149	42	92	90	53
O001_____	1.0015410	150	43	35	30	16
O002_____	0.9943007	151	44	26	18	17
O003_____	0.9978977	152	45	16	13	6
O004_____	0.9959553	153	46	14	18	7
P000_____	1.0011506	159	47	185	162	104
S000_____	1.0019473	161	48	178	155	97
N_____	0.9986045	165	49	87	88	50
O_____	0.9998081	166	50	259	212	132
S_____	1.0064372	168	51	18	26	11
[_____	0.9972849	170	52	64	74	46
c000_____	1.0119811	171	53	60	45	31
c006_____	1.0054158	176	54	64	55	35
c012_____	1.0054985	179	55	36	49	21
c_____	0.9989636	188	56	1236	1215	684
(___(_____	1.0059636	192	57	61	79	29
1___(_____	1.0018887	193	58	115	119	68
=___(_____	0.9963131	202	59	79	59	40
C___(_____	1.0039035	206	60	494	399	268
C___1_____	0.9991840	207	61	59	52	38
C___2_____	1.0020723	208	62	15	10	19
C___=_____	0.9941708	210	63	60	37	33
C___C_____	0.9994283	211	64	320	191	134
Br__(_____	1.0011776	212	65	29	44	14
Cl__(_____	0.9992838	215	66	229	226	106
Cl__1_____	1.0030256	216	67	17	12	9
Cl__C_____	0.9999034	219	68	20	23	10
F___(_____	1.0008248	220	69	42	111	5
N___(_____	1.0034944	223	70	52	50	28
N___+_____	1.0050316	224	71	15	15	11
N___C_____	0.9965020	228	72	21	14	17
O___(_____	1.0036181	230	73	201	148	88
O___−_____	1.0054058	231	74	15	15	12
O___=_____	0.9947318	235	75	103	70	54

Table 3 (*continued*)

SA_k	$CW(SA_k)$	ID	No.	NS_{TRN}	NS_{CLB}	NS_{TST}
O___C_____	1.0025416	236	76	68	61	32
S___(_____	1.0019934	242	77	21	21	14
[___(_____	1.0033912	249	78	41	51	33
[___+_____	0.9921431	250	79	15	19	11
[___−_____	1.0039739	251	80	15	15	12
[___N_____	1.0041361	256	81	15	15	11
[___O_____	1.0067672	257	82	15	17	12
c___(_____	1.0048399	261	83	553	496	284
c___1_____	1.0011618	262	84	554	535	294
c___2_____	0.9983495	263	85	126	117	63
c___3_____	1.0023584	264	86	23	29	20
c___C_____	1.0042609	268	87	20	18	16
c___Cl_____	1.0026704	270	88	32	29	14
c___N_____	1.0002041	271	89	21	21	11
c___O_____	0.9982729	272	90	24	33	15
c___c_____	1.0038082	275	91	540	558	318
(___C___(___	1.0006424	290	92	116	93	74
(___Cl__(___	1.0048728	293	93	81	82	38
(___F___(___	0.9957463	294	94	19	54	2
(___O___(___	0.9955458	297	95	25	17	12
(___c___(___	0.9986235	299	96	39	35	16
+___[___(___	0.9981487	300	97	15	17	11
1___C___(___	0.9957030	306	98	21	14	10
1___Cl__(___	0.9985849	309	99	17	11	9
1___c___(___	0.9987253	314	100	166	147	88
2___c___(___	1.0031084	327	101	40	29	16
2___c___1___	1.0058596	328	102	26	26	11
=___C___(___	1.0010580	344	103	23	17	10
=___O___(___	1.0015572	352	104	73	50	35
C___(___C___	0.9976632	358	105	135	110	79
C___(___1___	0.9940980	359	106	27	22	23
C___(___=___	1.0008726	360	107	42	24	20
C___(___(___	0.9963238	361	108	42	36	25
C___=___C___	1.0041450	376	109	16	10	10
C___C___(___	0.9994975	379	110	121	95	45
C___C___1___	1.0034394	380	111	19	20	13
C___C___C___	1.0003984	382	112	182	92	71
C___C___=___	1.0041736	383	113	17	6	9
C___O___(___	0.9968556	390	114	38	36	8
C___c___1___	0.9981276	396	115	20	18	16
Cl__(___Cl__	0.9986110	405	116	59	61	33
Cl__(___C___	1.0003258	408	117	14	20	4
Cl__(___(___	0.9946027	409	118	15	15	7
Cl__(___1___	1.0077613	411	119	46	36	11
Cl__C___(___	0.9976521	420	120	15	12	5
Cl__c___1___	1.0077941	421	121	32	29	14
F___(___(___	1.0015489	425	122	17	52	2
N___(___C___	0.9938391	430	123	20	11	5
N___c___1___	1.0052252	455	124	20	18	10
O___(___C___	0.9935385	460	125	41	22	13
O___(___O___	0.9928473	461	126	26	17	15
O___(___Cl__	0.9965902	462	127	17	5	4
O___(___(___	0.9975164	463	128	27	31	11
O___=___(___	0.9986326	471	129	71	50	35
O___=___C___	0.9998116	473	130	18	7	7
O___C___C___	0.9981001	476	131	39	27	13
O___C___(___	0.9972129	480	132	24	24	15
O___[___(___	0.9975276	487	133	14	14	12
O___c___1___	1.0027916	488	134	24	33	15
S___(___=___	1.0058827	494	135	14	12	8
[___+___N___	0.9993810	525	136	15	15	11
[___−___O___	1.0030368	526	137	15	15	12
[___N___+___	0.9953277	529	138	15	15	11
[___O___−___	1.0011927	531	139	15	15	12
c___(___C___	0.9982311	544	140	22	17	9
c___(___1___	0.9994668	548	141	14	14	9
c___(___c___	0.9994933	549	142	235	207	117
c___(___O___	0.9999477	550	143	16	10	4
c___1___c___	1.0004825	551	144	193	186	99
c___1___Cl__	1.0040200	553	145	17	12	8
c___1___C___	1.0044203	558	146	14	10	7
c___1___(___	0.9995601	560	147	102	105	58
c___2___c___	1.0044031	565	148	48	44	25
c___c___(___	0.9976437	597	149	264	240	139

(*continued on next page*)

Table 3 (*continued*)

SA_k	$CW(SA_k)$	ID	No.	NS_{TRN}	NS_{CLB}	NS_{TST}
c___c___2___	1.0030700	599	150	53	50	31
c___c___c___	0.9984242	600	151	257	283	158
c___c___1___	1.0005027	601	152	227	238	134

together with the calibration set (i.e., for the [training → test] scheme). The standard error of estimation(s) and the Fischer *F*-ratio are calculated after completion of the Monte Carlo optimization.

3. Results and discussion

Table 2 contains results of the modeling with **LimN** from 10 to 20. One can see from Table 2 that (1) the correlation balance gives better prediction for the external test sets for three random splits into a subtraining, a calibration, and a test sets; (2) the preferable results take place with different **LimN** values (17 for the split A [subtraining → calibration → test] scheme and 16 for the split A [training → test] scheme; 14 for the split B [subtraining → calibration → test] scheme and 19 for the split B [training → test] scheme; 17 for the split C [subtraining → calibration → test] scheme and 17 for the split C [training → test] scheme). Graphically this information is shown in Figs. 1–3 for the splits A, B, and C, respectively.

The best model obtained for split B with the balance of correlations (i.e., the [subtraining → calibration → test] scheme) is the following:

$$\log\ BCF = -71.0448\ (\pm 0.1775) + 67.1911\ (\pm 0.1631)DCW(14) \quad (3)$$

$n = 193$, $r^2 = 0.771$, $Q^2 = 0.767$, $s = 0.659$, $F = 643$ (subtraining set)

$n = 175$, $r^2 = 0.819$, $R^2_{pred} = 0.815$, $s = 0.578$, $F = 786$ (calibration set)

$n = 105$, $r^2 = 0.805$, $R^2_{pred} = 0.797$, $s = 0.528$, $F = 427$ (test set)-

where r is the correlation coefficient; s is the standard error of estimation; F is the Fischer *F*-ratio;

$$Q^2 = 1 - \frac{\sum\left[Y_{pred} - Y\right]^2}{\sum[Y - \overline{Y}(\text{training})]^2}\left(Y \text{ and } Y_{pred} \text{ on the training set}\right)$$

$$R^2_{pred} = 1 - \frac{\sum\left[Y_{pred} - Y\right]^2}{\sum[Y - \overline{Y}(\text{training})]^2}\left(Y \text{ and } Y_{pred} \text{ on the calibration or the test set}\right)$$

where Y and Y_{pred} are experimental and predicted values of the **log BCF**, respectively; $\overline{Y}$(training) is an average of the experimental values of the **log BCF** over the training set.

In brackets, the dispersion of the coefficients of regression in the leave-one-out calculations for the subtraining set is represented.

It should be noted that both the r and s must be taken into account for adequate estimation of the predictability of the QSPR/QSAR models [29], because it is a possible situation where a good r (together with the good F!) is accompanied by a poor s [29].

Recently, Roy and Roy [30] have suggested a criterion, denoted as R^2_m. For a model with good external predictability, R^2_m value should be greater than 0.5. In the case of the model calculated with Eq. (3), the test set is characterized by $R^2_m = 0.657$. Thus, the model is satisfactory, according to this criterion.

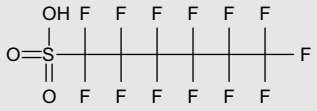
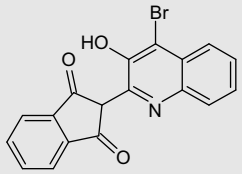
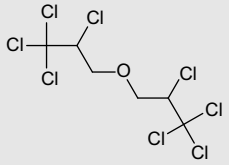
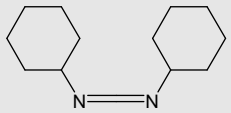
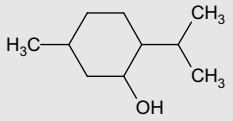
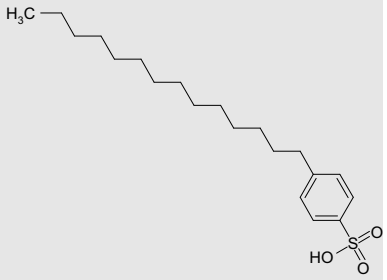
Table 4
The **DCW**(**14**) calculation: SMILES = ″CCCCC(C)C(C)C″; CAS = 3074-71-3; **DCW**(**14**) = 1.0990555.

SA_k	$CW(SA_k)$
C_____	1.0013833
C_____	1.0013833
C_____	1.0013833
C_____	1.0013833
C_____	1.0013833
(_____	0.9955638
C_____	1.0013833
(_____	0.9955638
C_____	1.0013833
(_____	0.9955638
C_____	1.0013833
(_____	0.9955638
C_____	1.0013833
C___C_____	0.9994283
C___C_____	0.9994283
C___C_____	0.9994283
C___C_____	0.9994283
C___(_____	1.0039035
C___(_____	1.0039035
C___(_____	1.0039035
C___(_____	1.0039035
C___(_____	1.0039035
C___(_____	1.0039035
C___(_____	1.0039035
C___(_____	1.0039035
C___C___C___	1.0003984
C___C___C___	1.0003984
C___C___C___	1.0003984
C___C___(___	0.9994975
C___(___C___	0.9976632
(___C___(___	1.0006424
C___(___C___	0.9976632
(___C___(___	1.0006424
C___(___C___	0.9976632
(___C___(___	1.0006424
C___(___C___	0.9976632
(002_____	0.9998299
=000_____	1.0061254
N000_____	1.0129657
O000_____	1.0088745
Cl00_____	1.0151967
Br00_____	0.9952189
F000_____	1.0020780
S000_____	1.0019473
P000_____	1.0011506
1000_____	1.0063275
2000_____	1.0055077
3000_____	1.0133248
4000_____	1.0014690
5000_____	0.9961318
C009_____	1.0
c000_____	1.0119811
!-09_____	1.0

Supplementary material section contains **log BCF** values of experimental and calculated with Eq. (3). Table 3 contains the list of active SMILES attributes (without of rare attributes) and their correlation weights used for the **DCW**(**14**) calculation. Table 4 contains an example of the **DCW**(**14**) calculation.

There are some compounds for which the prediction of the **log BCF** is not satisfactory if these substances fall into a test set. Table 5 shows the CAS numbers and structures of these compounds. Thus it is possible to identify at least five splits into a subtraining, a calibration and a test set for which the model will give unsatisfactory prediction due to the presence of an outlier. In spite of it, one can agree that the statistical characteristics of the best models for three splits indicate reasonable predictive potential of the SMILES-based optimal descriptors for QSPR modeling of **log BCF**.

Table 5
List of compounds which become outliers when placed into the test set.

CAS number	Structure
355-46-4	
10319-14-9	
127-90-2	
538-75-0	
2216-51-5	
47377-16-2	

A collection of statistical characteristics for the **log BCF** model (obtained by taking into account the octanol/water partition coefficient log P_{ow}) from [31] contains R^2 values ranged from 0.31 to 0.73 for non-ionic compounds ($n = 610$); from 0.19 to 0.62 for ionic compounds ($n = 84$); and from 0.32 to 0.74 for all compounds ($n = 694$). Thus, SMILES-based optimal descriptors obtained with the balance of correlations (without data on octanol/water partition coefficient) gave reasonable good prediction of the **log BCF** values for the external test sets.

List of the active SMILES attributes (Table 3) gives possibility to detect molecular fragments (encoded in SMILES) which are not rare for given subtraining set. In order to estimate significance of the attributes (molecular fragments) one can estimate the effect of blocking a given attribute for the model, for instance as a difference in correlation coefficient: before this blocking and after. Comparison of the numbers of the attributes for different splits (i.e., presence into the subtraining set, calibration set, and test sets) also can be a heuristic information. In particular this information can be useful for definition of the domain of applicability of the models. Authors are planning to carry out the analysis of these aspects of SMILES-based models in the near future.

It is a useful ability of the described approach that the SMILES attributes can be a representation not only of the molecular fragments but also they can be a representation of some situations in molecules which are not fragments (e.g., presence/absence of cycles; presence/absence of a chemical elements; presence/absence of double chemical bonds; presence/absence of rigid or vice versa of flexible regions in molecule, etc.). The behaviour of substances can be influenced by the above-mentioned situations. Hence, correlation weights of these situations can be a significant addition for well-known methods of the contributions of molecular fragments, such as the Free-Wilson approach [32], Fujita-Ban calculations [33], and semi-empirical topological indices [34].

It should be noted that the number of studies dedicated to QSPR/QSAR analyses with using SMILES is gradually increasing [35–39]. The existence of databases available via internet with elucidation of the molecular structure by SMILES is an apparent reason to search for SMILES-based QSPR/QSAR models.

4. Conclusions

1. SMILES-based optimal descriptors which are calculated by the Monte Carlo method give reasonable prediction of the bioconcentration factor for three splits into the subtraining set, calibration set, and external test set;
2. The correlation balance (i.e., the described [subtraining → calibration → test] scheme) is improving the **log BCF** model in comparison with the more traditional [training → test] scheme.

Acknowledgements

The authors thank the Marie Curie Fellowship (the contract ID 39036, CHEMPREDICT) and the EC project CAESAR (Project no. 022674 (SSPI)) for financial support.

Appendix. Supplementary material

Supplementary material associated with this article can be found in the online version, at doi: 10.1016/j.ejmech.2009.01.023.

References

[1] C.E. Cowan-Ellsberry, S.D. Dyer, S. Erhardt, M.J. Bernhard, A.L. Roe, M.E. Dowty, A.V. Weisbrod, Chemosphere 70 (2008) 1804–1817.
[2] S.D. Dimitrov, N.C. Dimitrova, J.D. Walker, G.D. Veith, O.G. Mekenyan, Pure Appl. Chem. 74 (2002) 1823–1830.
[3] http://www.epa.gov/.
[4] P.P. Roy, J.T. Leonard, K. Roy, Chemom. Intell. Lab. 90 (2008) 31–42.
[5] S.R. Johnson, J. Chem. Inf. Model. 48 (2008) 25–26.
[6] K. Roy, P.P. Roy, Chem. Biol. Drug Des. 71 (2008) 464–473.
[7] D. Weininger, J. Chem. Inf. Comput. Sci. 28 (1988) 31–36.
[8] D. Weininger, A. Weininger, J.L. Weininger, J. Chem. Inf. Comput. Sci. 29 (1989) 97–101.
[9] D. Weininger, J. Chem. Inf. Comput. Sci. 30 (1990) 237–243.
[10] http://www.daylight.com/.
[11] A.A. Toropov, B.F. Rasulev, J. Leszczynski, Bioorg. Med. Chem. 16 (2008) 5999–6008.
[12] S. Kumar, M. Kumar, K. Thurow, R. Stoll, U. Kragl, Environ. Model. Softw. 24 (2009) 44–53.
[13] C.-J. Feng, W.-H. Yang, L.-L. Mu, Jiegou Huaxue 27 (2008) 575–587.
[14] Y. Wang, Y. Li, J. Ding, Z. Jiang, Y. Chang, SAR QSAR Environ. Res. 19 (2008) 375–395.
[15] S. Cui, J. Yang, S. Liu, L. Wang, QSARs Sci. China Ser. B Chem. 50 (2007) 587–592.
[16] E. Papa, J.C. Dearden, P. Gramatica, Chemosphere 67 (2007) 351–358.

[17] K. Roy, I. Sanyal, P.P. Roy, SAR QSAR Environ. Res. 17 (2006) 563–582.
[18] T. Ivanciuc, O. Ivanciuc, D.J. Klein, Mol. Divers. 10 (2006) 133–145.
[19] W. Di Marzio, M.E. Saenz, Environ. Toxicol. 21 (2006) 118–124.
[20] P. Gramatica, E. Papa, QSAR Comb. Sci. 24 (2005) 953–960.
[21] J.M. Bermúdez-Saldaña, L. Escuder-Gilabert, M.J. Medina-Hernández, R.M. Villanueva-Camañas, S. Sagrado, J. Chromatogr. A 1063 (2005) 153–160.
[22] P.V. Khadikar, S. Singh, D. Mandloi, S. Joshi, A.V. Bajaj, Bioorg. Med. Chem. 11 (2003) 5045–5050.
[23] M.H. Fatemi, M. Jalali-Heravi, E. Konuze, Anal. Chim. Acta 486 (2003) 101–108.
[24] P. Gramatica, E. Papa, QSAR Comb. Sci. 22 (2003) 374–385.
[25] J.A. Arnot, F.A.P.C. Gobas, QSAR Comb. Sci. 22 (2003) 337–345.
[26] S.D. Dimitrov, N.C. Dimitrova, J.D. Walker, G.D. Veith, O.G. Mekenyan, QSAR Comb. Sci. 22 (2003) 58–68.
[27] C. Zhao, E. Boriani, A. Chana, A. Roncaglioni, E. Benfenati, Chemosphere 73 (2008) 1701–1707.
[28] A.A. Toropov, E. Benfenati, Bioorg. Med. Chem. 16 (2008) 4801–4809.
[29] A.A. Toropov, E. Benfenati, Curr. Drug Discov. Technol. 4 (2007) 77–116.
[30] P.P. Roy, K. Roy, QSAR Comb. Sci. 27 (2008) 302–313.
[31] M. Pavan, T.I. Netzeva, A.P. Worth, QSAR Comb. Sci. 27 (2008) 21–31.
[32] M.-O. Fouchécourt, M. Béliveau, K. Krishnan, Sci. Total Environ. 274 (2001) 125–135.
[33] V.K. Gombar, V.K. Kapoor, Eur. J. Med. Chem. 25 (1990) 689–695.
[34] B. da Silva Junkes, A.C.S. Arruda, R.A. Yunes, L.C. Porto, V.E.F. Heinzen, J. Mol. Model. 11 (2005) 128–134.
[35] D. Vidal, M. Thormann, M. Pons, J. Chem. Inf. Model. 45 (2005) 386–393.
[36] D. Vidal, M. Thormann, M. Pons, J. Chem. Inf. Model. 46 (2006) 836–843.
[37] D. Vidal, J. Blobel, Y. Pérez, M. Thormann, M. Pons, Eur. J. Med. Chem. 42 (2007) 1102–1108.
[38] G. Carta, V. Onnis, A.J.S. Knox, D. Fayne, D.G. Lloyd, J. Comput.-Aided Mol. Des. 20 (2006) 179–190.
[39] M. Olah, C. Bolog, T.I. Oprea, J. Comput.-Aided Mol. Des. 18 (2004) 437–449.