



The sequence of amino acids as the basis for the model of biological activity of peptides

Alla P. Toropova¹ · Maria Raškova² · Ivan Raška Jr.² · Andrey A. Toropov¹

Received: 22 June 2020 / Accepted: 15 December 2020 / Published online: 22 January 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE part of Springer Nature 2021

Abstract

The algorithm of building up a model for the biological activity of peptides as a mathematical function of a sequence of amino acids is suggested. The general scheme is the following: The total set of available data is distributed into the active training set, passive training set, calibration set, and validation set. The training (both active and passive) and calibration sets are a system of generation of a model of biological activity where each amino acid obtains special correlation weight. The numerical data on the correlation weights calculated by the Monte Carlo method using the CORAL software (<http://www.insilico.eu/coral>). The target function aimed to give the best result for the calibration set (not for the training set). The final checkup of the model is carried out with data on the validation set (peptides, which are not visible during the creation of the model). Described computational experiments confirm the ability of the approach to be a tool for the design of predictive models for the biological activity of peptides (expressed by pIC50).

Keywords QSAR · Amino acid · Peptide · Monte Carlo method · Index of ideality of correlation

1 Introduction

History of mathematical chemistry contains contributions of many outstanding scientists, such as A.T. Balaban, M. Randić, I. Gutman, N. Trinajstić, S.C. Basak, R. Carbó-Dorca, as well as many others [1–15]. Mathematical chemistry [1] is the area of research engaged in novel applications of mathematics to chemistry, biochemistry, and biology. It concerns itself principally with the mathematical modeling of complex molecular phenomena [2].

Most areas of research in mathematical chemistry include chemical graph theory, which deals with the development of topological descriptors which find application in quantitative

structure–property relationships [3, 4], as well as chemical aspects of group theory, which finds applications in stereochemistry and quantum chemistry [5, 6].

Cheminformatics is a relatively young field of natural sciences. By analogy with "in viva" and "in vitro," the results of cheminformatics denominate as "in silico" [7].

It is to be noted, contributions of Prof. R. Carbó-Dorca, related to the development of cheminformatics tools applied to quantum mechanical theoretical problems, which gave the possibility to solve chemical problems, like catalysis and reactivity, by simple computational schemes [8–12]. Cheminformatic gradually extends to solve tasks in fields of theoretical chemistry, computational chemistry, and modeling [13–15].

Apply mathematical methods to solve the tasks of chemistry and biochemistry can be effective [16, 17]. Peptides are important objects of chemistry, biochemistry, and medicine. Most interest in using proteins and peptides is caused by their application in drug design [18]. The amino acid residues of epitope-peptide substrate and SARS coronavirus main protease are interacting. Hence, the affinity of epitope-peptides with class I MHC (major histocompatibility complex) molecules can be used to development of antiviral agents, e.g., toward coronaviruses [18].

Published as part of the special collection of articles "Festschrift in honour of Prof. Ramon Carbó-Dorca".

✉ Andrey A. Toropov
andrey.toropov@marionegri.it

¹ Department of Environmental Health Science, Laboratory of Environmental Chemistry and Toxicology, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2, 20156 Milan, Italy

² 3Rd Medical Department, 1st Faculty of Medicine, Charles University in Prague, U Nemocnice 1, 12808 Prague 2, Czech Republic

A fundamentally widely accepted science principle to understand complex systems is “Everything should be made as simple as possible, but no simpler” [19]. Perhaps, the approach used here cannot be adequately evaluated using the above principle, since a simpler method is not possible, or at least a simpler approach has not yet been described in the literature [20–23]. To state the approach “simpler” than “simple” is not correct, since the approach gives quite good models [20–23]. The model of biological activity of peptides described here is based on sequences of amino acids, represented by 1-letter codes (Table 1).

The aim of the present study is the estimation of the CORAL software to provide a satisfactory model for the bioactivity of peptides. Representation of peptides via a sequence of amino acids is like a well-known simplified molecular input-line entry system (SMILES) [24]. Consequently, the CORAL software (www.insilico.eu/coral) that is oriented to build up quantitative structure–activity relationships (QSARs) using the SMILES representation can be a tool to build up a predictive model for the activity of peptides as a function of sequences of the 1-letter codes of corresponding amino acids [25]. Factually, the sequences of amino acids represented by 1-letter codes are quasi-SMILES [20, 21].

2 Method

2.1 Data

The numerical data on the biological activity of epitope-peptides with class I MHC (major histocompatibility complex) molecules taken from the literature [18]. The endpoint expressed via a negative logarithm of half-maximal inhibitory concentration IC_{50} (pIC_{50}). Table 1 contains sequences of amino acids represent epitope-peptides studied here.

The available epitope-peptides were randomly distributed into the active training set (25%), passive training set (25%), calibration set (25%), and validation set (25%). Each above set has a defined task. The task for the active training set is to build up optimal correlation weights for the optimal descriptor. The task for the passive training set is to checkup whether current correlation weights (and the optimal descriptor) are satisfactory for peptides, which are not involved in the calculation of the correlation weights. The task for the calibration set is to detect the moment of the begin overtraining. The task of peptides from the validation set is the final estimation of the predictive potential of the model.

Table 1 Structures and 1-letter codes for Amino acids

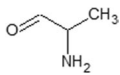
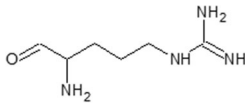
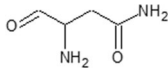
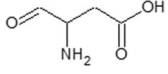
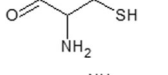
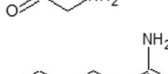
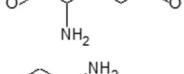
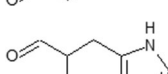
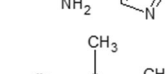
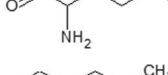
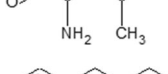
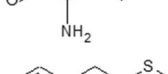
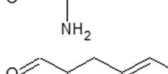
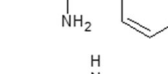
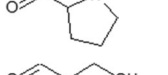
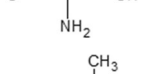
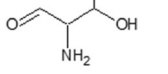
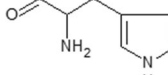
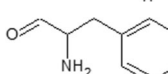
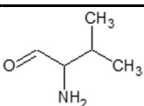
Amino acid	1-letter code	Structure
Alanine	A	
Arginine	R	
Asparagine	N	
Aspartic Acid	D	
Cysteine	C	
Glutamic acid	E	
Glutamine	Q	
Glycine	G	
Histidine	H	
Isoleucine	I	
Leucine	L	
Lysine	K	
Methionine	M	
Phenylalanine	F	
Proline	P	
Serine	S	
Threonine	T	
Tryptophan	W	
Tyrosine	Y	

Table 1 (continued)

Amino acid	1-letter code	Structure
Valine	V	

2.2 Quantitative structure–activity relationships (QSARs)

The CORAL software provides models, which are linear one-variable correlations obtained by the Monte Carlo method (<http://www.insilico.eu/coral>). The generalized representation of the model for the biological activity of peptides is the following one-variable correlation:

$$pIC_{50} = C_0 + C_1 \times DCW(T, N) \quad (1)$$

The $DCW(T, N)$ is the descriptor of correlation weights (DCW). The C_0 and C_1 are regression coefficients. The T and N are parameters of the Monte Carlo optimization discussed below.

2.3 The descriptor of correlation weights (DCW)

The descriptors applied to QSAR analysis are calculated as the following:

$$DCW(T^*, N^*) = \sum CW(A_k) \quad (2)$$

The A_k is a 1-letter code of amino acid; $CW(A_k)$ is the correlation weights for the A_k .

The T is an integer to define two classes (i) the rare and (ii) non-rare. If the frequency of A_k in the active training set is less than T , the A_k is rare, and the $CW(A_k) = 0$ (i.e., the A_k is removed from the modeling process). Thus, the model is based on correlation weights solely non-rare in the active training set amino acids. The N is the number of iterations for the Monte Carlo optimization. The $T = T^*$ and $N = N^*$ are values which provide the best statistical quality of the model for the calibration set.

2.4 Monte Carlo optimization

The scheme of the Monte Carlo optimization is described in [23, 25]. The essence of this version of the optimization procedure is the application of the Index of ideality of correlation (IIC). Models for the inhibitory activity of peptides built up here are build up to apply two different target functions TF_1 and TF_2 :

$$TF_1 = R_{AT} + R_{PT} - |R_{AT} - R_{PT}| * 0.1 \quad (3)$$

$$TF_2 = TF_1 + IIC_{CLB} * 0.5 \quad (4)$$

The R_{AT} and R_{PT} are the correlation coefficient between observed and predicted endpoints for the active training set and passive training set, respectively.

The IIC_{CLB} is calculated with data on the calibration set as the following:

$$IIC_{CLB} = r_{CLB} \frac{\min(-MAE_{CLB}, +MAE_{CLB})}{\max(-MAE_{CLB}, +MAE_{CLB})} \quad (5)$$

$$-MAE_{CLB} = \frac{1}{-N} \sum_{k=1}^{-N} |\Delta_k|, \Delta_k \leq 0; -N \text{ is the number of } \Delta_k \leq 0 \quad (6)$$

$$+MAE_{CLB} = \frac{1}{+N} \sum_{k=1}^{+N} |\Delta_k|, \Delta_k \leq 0; +N \text{ is the number of } \Delta_k \leq 0 \quad (7)$$

$$\Delta_k = observed_k - calculated_k \quad (8)$$

The observed and calculated are the corresponding values of pIC_{50} .

Figure 1 contains the comparison of histories of the Monte Carlo optimizations with target functions TF_1 and TF_2 . One can see, the TF_2 seems preferable because factually the decrease in the statistical quality for calibration set and validation set is not observed, whereas in the case of TF_1 the decrease in the statistical quality for the calibration set and validation set is observed.

2.5 Domain of applicability

The domain of applicability for the CORAL model is defined according to the distribution of SMILES attributes in the active training set and calibration set as two steps:

Step 1: the definition of the statistical defect (d_k) for each SMILES attribute involved in building up of a model:

$$d_k = \frac{|P(A_k) - P'(A_k)|}{N(A_k) + N'(A_k)} \quad (9)$$

where $P(A_k)$ and $P'(A_k)$ are the probability of A_k in the training and calibration sets, respectively.

$N(A_k)$ and $N'(A_k)$ are frequencies of A_k in the training and calibration sets, respectively.

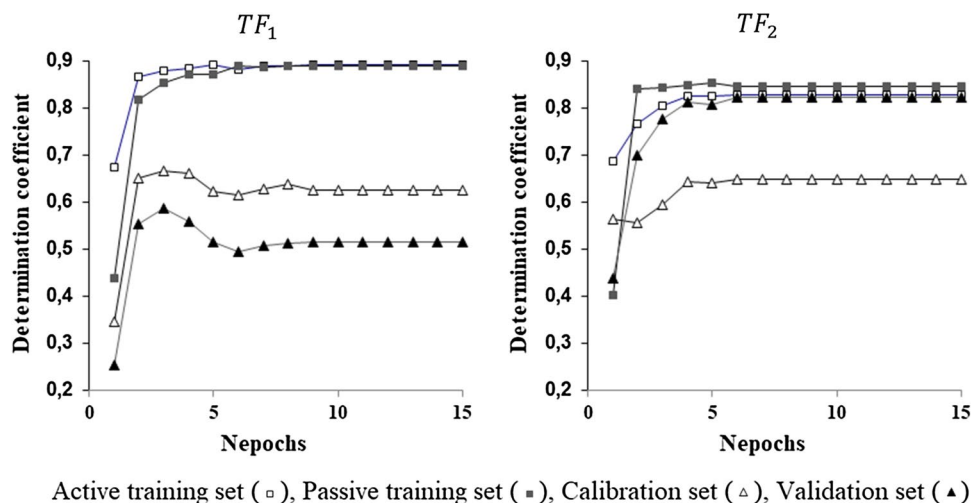
Step 2: the calculation for all substances the statistical SMILES-defect (D_j):

$$D_j = \sum_{k=1}^{NA} d_k \quad (10)$$

where NA is the number of non-blocked SMILES attributes in the SMILES.

A substance falls in the domain of applicability if

Fig. 1 Histories of the Monte Carlo optimization (Split 1) with target functions TF_1 and TF_2



$$D_j < 2 * \bar{D} \quad (11)$$

where \bar{D} is the average of the statistical SMILES-defect for the training set.

The same operation can be carried out with the sequences of 1-letter codes of amino acids, if instead of A_k defined as a SMILES attribute, one examined A_k defined as a 1-letter code of corresponding amino acids.

3 Results and discussion

The models obtained for three random splits into the training set (which is association of the active and passive training sets together with the calibration set) and validation set are the following:

Target Function TF_1

$$pIC_{50} = 5.0637012 (\pm 0.3150527) + 0.9790357 (\pm 0.1064904) * DCW(1, 3) \quad (12)$$

$$pIC_{50} = 5.3015843 (\pm 0.1783155) + 1.4109089 (\pm 0.1001528) * DCW(1, 3) \quad (13)$$

$$pIC_{50} = 2.6879582 (\pm 0.2459626) + 1.0011131 (\pm 0.0482456) * DCW(1, 3) \quad (14)$$

Target Function TF_2

$$pIC_{50} = 4.0179522 (\pm 0.5296001) + 0.4553542 (\pm 0.0634366) * DCW(1, 15) \quad (15)$$

$$pIC_{50} = 4.8689021 (\pm 0.3087049) + 0.6850851 (\pm 0.0712025) * DCW(1, 15) \quad (16)$$

$$pIC_{50} = 5.3828941 (\pm 0.4250702) + 0.7649124 (\pm 0.1215301) * DCW(1, 15) \quad (17)$$

Table 2 contains the statistical characteristics of the models calculated with Eqs. 12–17.

One can see, the predictive potential of models calculated using the IIC is better.

Having numerical data on correlation weights of different amino acids obtained in several runs of the optimization, one can detect the amino acids of two classes: (1) amino acids with stable positive correlation weights, these are promoters of increase of pIC_{50} ; and (2) amino acids with stable negative correlation weights, these are promoters of decrease of pIC_{50} . Thus, the approach gives the statistical mechanistic interpretation of the models. Table 3 contains a collection of amino acids which are promoters of increase/decrease for pIC_{50} . It is to be noted, the prevalence of corresponding amino acids also should be considering.

Table 4 contains experimental and calculated with Eq. 17 pIC_{50} . Table 5 contains the numerical data on the correlation

weights of amino acids to calculate the model with Eq. 17.

Table 2 Statistical quality of models for three random splits

Split	Set	R^2	Q^2	IIC	RMSE
Optimization with TF_1					
1	Active training	0.7625	0.5558	0.8732	0.360
	Passive training	0.8250	0.7065	0.6739	0.395
	Calibration	0.6012	0.4017	0.3695	0.506
	Validation	0.6220	0.4816		0.490
2	Active training	0.8205	0.7052	0.9058	0.333
	Passive training	0.9165	0.8301	0.4709	0.374
	Calibration	0.5223	0.2836	0.4258	0.592
	Validation	0.5481	0.3476		0.515
3	Active training	0.8846	0.8229	0.9406	0.265
	Passive training	0.7283	0.5982	0.8264	0.599
	Calibration	0.5053	0.2612	0.3745	0.927
	Validation	0.5900	0.3277		0.700
Optimization with TF_2					
1	Active training	0.6416	0.3506	0.5340	0.442
	Passive training	0.7231	0.5868	0.4120	0.507
	Calibration	0.9486	0.9157	0.9679	0.142
	Validation	0.7766	0.6298		0.306
2	Active training	0.6976	0.4905	0.5568	0.432
	Passive training	0.9543	0.9192	0.8516	0.332
	Calibration	0.7102	0.5447	0.8406	0.337
	Validation	0.7856	0.6596		0.270
3	Active training	0.5326	0.1846	0.7298	0.533
	Passive training	0.8128	0.6796	0.6251	0.562
	Calibration	0.8743	0.8139	0.8827	0.214
	Validation	0.7909	0.6721		0.248

Each set contains ten peptides

The best model is indicated by bold

Table 3 Amino acids which are promoters of increase / decrease for pIC₅₀ for examined peptides

Comment	A_k	CWsProbe 1	CWsProbe 2	CWsProbe 3	N_{AT}	N_{PT}	N_C	d_k
Increase	V.....	0.47695	0.30991	0.26611	10	10	10	0.0000
	L.....	1.29542	0.73164	0.31587	8	5	7	0.0067
	F.....	1.07326	0.70770	0.37614	6	6	7	0.0077
	I.....	0.76211	0.16717	0.34684	6	3	4	0.0200
	A.....	0.54686	0.01821	0.06304	4	3	2	0.0333
	G.....	0.44966	0.52819	0.73395	4	5	4	0.0000
	Y.....	1.46411	0.65332	0.40546	4	5	5	0.0111
	M.....	1.29967	0.55126	0.39601	2	0	3	0.0200
Decrease	T.....	-0.26044	-0.28480	-0.34702	6	9	6	0.0000
	E.....	-0.62472	-0.62778	-0.55954	1	3	1	0.0000

N_{AT} , N_{PT} , and N_C are the frequencies of an amino acid in the active training set, passive training set, and the calibration set, respectively

Table 6 contains an example of calculation DCW(1,15) for epitope-peptide “WLEPGPVTA” together with the calculation of corresponding pIC₅₀ using Eq. 17.

Thus, the described approach can be a tool to build up models for pIC₅₀ for epitope-peptides.

Table 4 Experimental and calculated with Eq. 17 pIC_{50} for model obtained with split 3 (the best model): “+” is the indicator for the active training set; “-” is the indicator for the passive training set; “#” is the indicator of calibration set; and “*” is the indicator for validation set

Set	ID	Sequence of amino acids	$DCW(1,15)$	pIC_{50} Expr	pIC_{50} Calc	$D_f(\bar{D} = 0.08757)$	Applicability
-	P01	WLEPGPVTA	1.98966	6.0820	6.9048	0.0754	YES
-	P02	ITSQVPFSV	1.62921	6.1960	6.6291	0.1259	YES
#	P03	FLEPGPVTA	2.17966	6.8980	7.0501	0.0485	YES
#	P04	ITAQVPFSV	2.21389	7.0200	7.0763	0.1029	YES
+	P05	YLEPGPVTL	2.98174	7.0580	7.6637	0.0421	YES
#	P06	YTDQVPFSV	2.39417	7.0660	7.2142	0.0862	YES
-	P07	YLEPGPVTI	2.21031	7.1870	7.0736	0.0754	YES
*	P08	YLEPGPVTV	2.20698	7.3420	7.0710	0.0421	YES
#	P09	YLSPGPVTA	3.06834	7.3830	7.7299	0.0651	YES
#	P10	IIDQVPFSV	3.11987	7.3980	7.7693	0.1219	YES
+	P11	ITWQVPFSV	1.93195	7.4630	6.8607	0.1529	YES
+	P12	ITYQVPFSV	2.14528	7.4800	7.0238	0.1195	YES
#	P13	ILSQVPFSV	3.05039	7.6990	7.7162	0.1117	YES
-	P14	IMDQVPFSV	2.69191	7.7190	7.4420	0.0886	YES
*	P15	YLMPGPVTV	3.23638	7.9320	7.8584	0.0421	YES
#	P16	WLDQVPFSV	3.60203	7.9390	8.1381	0.1052	YES
*	P17	YLAPGPVTA	3.65302	8.0320	8.1771	0.0421	YES
+	P18	YLYPGPVTV	3.58840	8.0510	8.1277	0.0587	YES
*	P19	YLWPGPVTV	3.37507	8.1250	7.9645	0.0921	YES
#	P20	ILYQVPFSV	3.56646	8.3100	8.1109	0.1052	YES
-	P21	ILDQVPFSV	3.89130	8.4810	8.3594	0.0886	YES
-	P22	YLFPGPVTA	3.56108	8.4950	8.1068	0.0651	YES
+	P23	YLDQVPFSV	3.81535	8.6380	8.3013	0.0719	YES
-	P24	ILFQVPFSV	3.54314	8.6990	8.0931	0.1117	YES
-	P25	ILWQVPFSV	3.35313	8.7700	7.9477	0.1386	YES
+	P26	WTDQVPFSV	2.18084	6.1450	7.0510	0.1195	YES
*	P27	YLEPGPVTA	2.20298	6.6680	7.0680	0.0421	YES
*	P28	ITDQVPFSV	2.47011	6.9470	7.2723	0.1029	YES
*	P29	ITFQVPFSV	2.12196	7.1790	7.0060	0.1259	YES
*	P30	FTDQVPFSV	2.37085	7.2120	7.1964	0.0926	YES
-	P31	ITMQVPFSV	1.79326	7.3980	6.7546	0.1029	YES
#	P32	YLSPGPVTV	3.07233	7.6420	7.7330	0.0651	YES
+	P33	YLYPGPVTA	3.58440	7.7720	8.1246	0.0587	YES
+	P34	YLAPGPVTV	3.65702	7.8180	8.1802	0.0421	YES
*	P35	ILAQVPFSV	3.63508	7.9390	8.1634	0.0886	YES
*	P36	ILMQVPFSV	3.21445	8.1250	7.8417	0.0886	YES
#	P37	YLFPGPVTV	3.56508	8.2370	8.1099	0.0651	YES
-	P38	YLMPGPVTA	3.23239	8.3670	7.8554	0.0421	YES
+	P39	YLWPGPVTA	3.37107	8.4950	7.9615	0.0921	YES
+	P40	FLDQVPFSV	3.79203	8.6580	8.2835	0.0783	YES

4 Conclusions

The described approach gives a robust model for the biological activity of peptides (Table 4). The results are quite acceptable for three random splits into the training set and validation set. The approach obeys the OECD principles [26]. Once again, the possibility to build up predictive

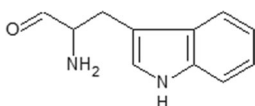
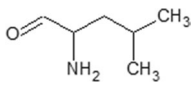
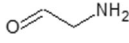
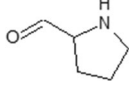

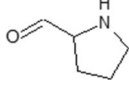
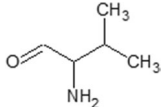
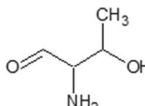
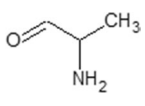
models for endpoints related to complex molecular systems (peptides) is confirmed [5–8]. In addition, the described confirms once more that applying the *IIC* improves the predictive potential of models [20, 25].

Table 5 Numerical data on the correlation weights to calculate model with Eq. 17

Amino acids, A_k	$CW(A_k)$	N_{AT}	N_{PT}	N_C	d_k
A.....	0.42063	3	3	3	0.0000
D.....	0.67685	3	2	3	0.0000
E.....	- 1.02940	1	2	1	0.0000
F.....	0.32870	5	7	8	0.0231
G.....	0.17820	5	4	4	0.0111
I.....	0.42796	2	7	4	0.0333
L.....	1.19938	7	7	7	0.0000
M.....	0.0	0	3	0	0.0000
P.....	0.43966	10	10	10	0.0000
Q.....	0.13354	5	6	6	0.0091
S.....	- 0.16405	5	6	8	0.0231
T.....	- 0.22180	8	6	6	0.0143
V.....	0.42463	10	10	10	0.0000
W.....	0.13869	3	2	1	0.0500
Y.....	0.35202	7	3	5	0.0167

N_{AT} , N_{PT} , and N_C are the frequencies of an amino acid in the active training set, passive training set, and the calibration set, respectively

Table 6 Calculation of $DCW(1,15)$ and pIC_{50} for epitope-peptide = WLEPGPVTA

A_k	Structure	$CW(A_k)$
W		0.13869
L		1.19938
E		- 1.02940
P		0.43966
G		0.17820
P		0.43966
V		0.42463
T		- 0.22180
A		0.42063
$DCW(1,15) = \sum CW(A_k)$		1.98966
$pIC_{50} = 5.3828941 + 0.7649124 * DCW(1,15)$		6.9048

Acknowledgements AAT and APT are grateful for the contribution of the project LIFE-VERMEER contract (LIFE16 ENV/ES/000167) for financial support.

Compliance with ethical standards

Conflict of interest The authors confirm they have no conflict of interest.

References

- Restrepo G (2016) Mathematical chemistry, a new discipline. In: Scerri E, Fisher G (Eds) Essays in the philosophy of chemistry. Oxford University Press, New York, UK, Chapter 15, pp 332–351.
- Gutman I, Polansky OE (1988) Mathematical concepts in organic chemistry. SIAM Review 30(2):348–350
- Trinajstić N, Gutman I (2002) Mathematical chemistry. Croat Chem Acta 75:329–356
- Balaban AT (2005) Reflections about mathematical chemistry. Found Chem 7:289–306
- Restrepo G, Villaveces JL (2012) Mathematical thinking in chemistry. HYLE 18:3–22
- Basak SC, Restrepo G, Villaveces JL (Eds) (2015) Advances in mathematical chemistry and applications. Volume 2. Bentham Science eBooks. ISBN: 9781681080529
- Engel T (2006) Basic overview of chemoinformatics. J Chem Inf Model 46(6):2267–2277
- Fradera X, Amat L, Besalú E, Carbó-Dorca R (1997) Application of molecular quantum similarity to QSAR. Quant Struct-Act Rel 16(1):25–32
- Carbó-Dorca R (2007) About the prediction of molecular properties using the fundamental quantum QSPR (QQSPR) equation. SAR QSAR Environ Res 18(3–4):265–284
- Poater A, Saliner AG, Carbó-Dorca R, Poater J, Solà M, Cavallo L, Worth AP (2009) Modeling the structure-property relationships of nanoneedles: a journey toward nanomedicine. J Comput Chem 30(2):275–284
- Carbó-Dorca R, Besalú E (2011) Construction of coherent nano quantitative structure-properties relationships (nano-QSPR) models and catastrophe theory. SAR QSAR Environ Res 22(7–8):661–665
- Ayers PL, Boyd RJ, Bultinck P, Caffarel M, Carbó-Dorca R, Causá M, Cioslowski J, Contreras-García J, Cooper DL, Coppens P, Gatti C, Grabowsky S, Lazzeretti P, Macchi P, Martín Pendás Á, Popelier PLA, Ruedenberg K, Rzepa H, Savin A, Sax A, Schwarz WHE, Shahbazian S, Silvi B, Solà M, Tsirelson V (2015) Six questions on topology in theoretical chemistry. Comput Theor Chem 1053:2–16
- Carbó-Dorca R (2018) Toward a universal quantum QSPR operator. Int J Quantum Chem 118(15):1
- Carbó-Dorca R, Chakraborty T (2019) Divagations about the periodic table: BOOLEAN hypercube and quantum similarity connections. J Comput Chem 40(30):2653–2663
- Carbó-Dorca R, Chakraborty T (2019) Hypercubes defined on n-ary sets, the Erdős–Faber–Lovász conjecture on graph coloring, and the description spaces of polypeptides and RNA. J Math Chem 57(10):2182–2194
- Carbó-Dorca R, Van Damme S (2007) Solutions to the quantum QSPR problem in molecular spaces. Theor Chem Acc 118(3):673–679
- Ponec R, Bultinck P, Van Damme S, Carbó-Dorca R, Tantillo DJ (2005) Geometric and electronic similarities between transition

- structures for electrocyclizations and sigmatropic hydrogen shifts. *Theor Chem Acc* 113(4):205–211
18. Du Q-S, Huang R-B, Wei Y-T, Wang C-H, Chou K-C (2007) Peptide reagent design based on physical and chemical properties of amino acid residues. *J Comput Chem* 28(12):2043–2050
 19. Hogeweg P (2010) Multilevel cellular automata as a tool for studying bioinformatic processes. In: Kroc J, Sloot P, Hoekstra A (eds) *Simulating complex systems by cellular automata. Understanding Complex Systems*, Springer, Berlin, Heidelberg, pp 19–28
 20. Toropov AA, Toropova AP, Leszczynska D, Leszczynski J (2019) “Ideal correlations” for biological activity of peptides. *Biosystems* 181:51–57
 21. Toropova AP, Toropov AA, Benfenati E, Leszczynska D, Leszczynski J (2018) Prediction of antimicrobial activity of large pool of peptides using quasi-SMILES. *Biosystems* 169–170:5–12
 22. Toropova AP, Toropov AA, Beeg M, Gobbi M, Salmona M (2017) Utilization of the monte carlo method to build up QSAR models for hemolysis and cytotoxicity of antimicrobial peptides. *Curr Drug Discov Technol* 14(4):229–243
 23. Toropov AA, Toropova AP, Raska I Jr, Benfenati E, Gini G (2012) QSAR modeling of endpoints for peptides which is based on representation of the molecular structure by a sequence of amino acids. *Struct Chem* 23(6):1891–1904
 24. Weininger D (1988) SMILES, a chemical language and information system: 1: Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 28(1):31–36
 25. Toropov AA, Carbó-Dorca R, Toropova AP (2018) Index of Ideality of correlation: new possibilities to validate QSAR: a case study. *Struct Chem* 29(1):33–38
 26. Toropova AP, Toropov AA (2014) CORAL software: prediction of carcinogenicity of drugs by means of the monte carlo method. *Eur J Pharm Sci* 52(1):21–25

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.