SciForum
Mol2Net

CORAL: The Dispersion of SWNTs in Different Organic Solvents

Alla P. Toropova, Andrey A. Toropov*
IRCCS, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, 20156, Via La Masa 19, Milano, Italy
*To whom correspondence should be addressed: E-mail: andrey.toropov@marionegri.it
Tel: +39 02 3901 4595 Fax: +3902 3901 4735 (AAT)。

Published: 4 December 2015

Abstract: Single-walled carbon nanotubes (SWNTs) are group of new substances with specific cylindrical architecture of their molecules. The dispersion of SWNTs in different organic solvents is parameter that can be valuable information for development of nanomaterials. The CORAL software is a tool to build up model for different endpoints using the Monte Carlo technique. In this work, the ability of the CORAL software to be a tool to predict dispersion of SWCTs in different organic solvents demonstrated.

Keywords: QSPR; Monte Carlo method; SMILES; Validation; Domain of applicability; CORAL software

1. Introduction

The development of nanotechnology indicates that use of carbon nanotubes (CNTs), in general, and single-walled nanotubes (SWNTs), in particular, gives attractive possibilities for chemical technology [1], biochemistry [2], and medicine [3]. The dispersibility of SWNTs in various solvents is important physicochemical characteristics [4] from point of view of technology [5, 6].

The theoretical approaches to predict of the endpoint for different solvents developed and described in the literature [5, 6]. Apparently, however, similar studies based on the quantitative structure – property / activity relationships (QSPRs/QSARs) [7-10] be continued.

In particular, this work dedicated to search for a new alternative approaches to predict the dispersibility of SWNTs in organic solvents using the Monte Carlo method [11, 12].

2. Method

2.1. Data

The dispersibility of SWNTs in a series of 29 different organic solvents taken in the literature [5, 6]. The endpoint is decimal logarithm of dispersibility C_{max} expressed in mg/mL. Three random splits into the visible

training set (in fact this is structured into two sets: the training and calibration sets) and the invisible validation set are examined in order to check up the actual ability of the approach.

2.2. Optimal descriptors

The optimal descriptor used in this work calculated as the following:

$$DCW(T^*, N^*) = \Sigma CW(V_k) \quad (1)$$

In Eq. 1: The T* is the coefficient to classify vertex degree into two categories rare and not rare. The parameter has influence upon the results of the Monte Carlo optimization

The Vk is vertex in the hydrogen-suppressed molecular graph [13-15]. Table 1 contains example of the hydrogen suppressed graph together with (0, 1) adjacency matrix and Vk values, which are calculated using the elements of the matrix; the $CW(V_k)$ is correlation weight of the V_k. The T* is threshold or a coefficient for the classification of vertices into two classes: (i) rare (the number of V_k in the training set is less than T^*) and (ii) active (the number of V_k in the training set is larger than T^*). The rare vertices are not involved building up model: their correlation weights fixed equal to zero. The N^* is the number of epochs of the Monte Carlo optimization. In fact, one can use arbitrary T and N, but the T* and N^* are values of these parameters which give preferable statistical quality of the model for the calibration set, hoping that the model is avoided of the overtraining (i.e. the situation where the excellent quality for the training set accompanied by poor quality for the calibration set).

[Table 1, around here]

Having the numerical data on the correlation weights, one can calculated the DCW(T*,N*) for all compounds of the training, calibration, and test sets. Using the data on the training set, one should to calculate the model

$$Endpoint = C_0 + C_1 * DCW(T^*,N^*) \quad (2)$$

The predictive potential of the model calculated with Eq. 2 should be checked with data on the calibration and validation sets.

2.3. Mechanistic interpretation

The CORAL models give the possibility to interpret the role of different molecular features as the promoters of increase or decrease of an endpoint. For instance, if in several runs of the Monte Carlo optimization the correlation weight of the V_k is larger than zero, then this feature is promoter of the endpoint increase, whereas if the correlation weights of the V_k are less than zero in several runs of the optimization then the V_k should be interpreted as promoter of the endpoint decrease.

2.4. Domain of applicability

The domain of applicability for the CORAL model defined according to prevalence of different molecular features in the training and the calibration sets: each molecular feature has the statistical defect. The defect is equal to difference between probabilities of the molecular feature in the training set and in the calibration set.

Ideal situation if the difference is zero, however in praxis, this value is not zero. Apparently, the preferable distribution should be characterized by the minimal sum of these parameters for all active molecular features. Thus, the approach gives possibility not only to define the domain of applicability, but, also, to compare different distributions into the training and calibration sets.

3. Results and Discussion

3.1. Models

The models for dispersibility of SWNTs in different organic solvents for three different random splits into the training, calibration, and validation sets are the following:

Split 1: $\log_{10}C_{max}$ = -2.9944 (± 0.1266) + 0.2076 (± 0.0183) * DCW(3,24) (3)

Split 2: $\log_{10}C_{max}$ = -3.2119 (± 0.1310) + 0.2380 (± 0.0200) * DCW(1,25) (4)

Split 3: $\log_{10}C_{max}$ = -3.1077 (± 0.1187) + 0.2165 (± 0.0175) * DCW(2,25) (5)

Table 2 contains numerical data on the correlation weights used to calculate the DCW(T*,N*) for calculation with Eqs. 3-5. Table 3 contains the statistical characteristics of models calculated with Eqs. 3-5.

[Table 2 and 3, around here]

3.2. Domain of applicability

The estimation of the domain has been done by scheme described in the literature [16]: the solvent with sum of defects for the SMILES less than average value of this parameter (for the training set) multiplied by 2:

$$\sum Defect \le 2 \times \overline{\sum defect} \quad (6)$$

[Table 4, around here]

One can see (Table 4) the distribution into the training, calibration, and validation sets has influence upon the domain of applicability, but this situation gives possibility to select preferable from the statistical point of view the distribution (minimum of the above-mentioned defect).

3.3. Mechanistic interpretation

Three runs of the Monte Carlo optimization with selected T* and N* give correlation weights collected in Table 5. One can hypothesizes about the role of molecular features represented by the V_k in the behavior of a solvent: if all runs give positive value of correlation weight for a V_k then the molecular feature can be classified as promoter of an endpoint increase, if all runs gives negative value of correlation weight then the molecular feature represented by the V_k can be classified as promoter of endpoint decrease [16].

3.4. Selection of molecular features for increase (decrease) of dispersibility of SWNT

The analysis of data collected in Table 5 lead to hypothesis that presence (in hydrogen suppressed molecular graph which is representation of a solvent) of carbon and nitrogen atoms with vertex degree 3, oxygen with vertex degree 1, and carbon atom with vertex degree 2 are promoter of dispersibility increase. The presence in molecular graph represented a solvent carbon vertex with vertex degree 1 is promoter of the endpoint decrease.

3.5. Comparison with QSAR models from the literature

The statistical characteristics of model of $\log_{10}C_{max}$ (for validation set, the same 29 solvents) suggested in work [5] are n=6, r^2=0.932; $\overline{r_m^2}$ =0.844, Δr_m^2 = 0.066; the statistical quality of model (for the same 29 solvents) suggested in work [6] are n=7, r^2=0.807; $\overline{r_m^2}$ =0.744, Δr_m^2 = 0.125. The above-mentioned models related to fixed splits into the

training and validation sets, whereas models suggested in this work are checked up with three different splits. It is to be noted, different splits into the training and validation sets used in work [5] and in work [6].

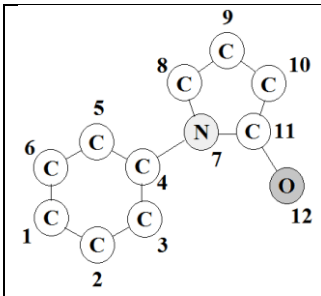
4. Conclusions

The described version of the Monte Carlo method gives satisfactory prediction for the disprsibility of SWNT in different solvents. The distribution into the visible training set (together with calibration set) and the invisible validation set has influence on the predictive potential models. The approach gives quite convenient measure of quality of distribution into the training and the validation sets together with convenient criterion of the domain of applicability.

Acknowledgements

Authors thank the EU project PROSIL (project LIFE12 ENV/ IT/ 000154) for financial support.

Table 1
Example of the hydrogen suppresed graph together with the adjacecncy matrix and vertex degree values (V_k).



	C_1	C_2	C_3	C_4	C_5	C_6	N_7	C_8	C_9	C_{10}	C_{11}	O_{12}	V_k
C_1	0	1	0	0	0	1	0	0	0	0	0	0	2
C_2	1	0	1	0	0	0	0	0	0	0	0	0	2
C_3	0	1	0	1	0	0	0	0	0	0	0	0	2
C_4	0	0	1	0	1	0	1	0	0	0	0	0	3
C_5	0	0	0	1	0	1	0	0	0	0	0	0	2
C_6	1	0	0	0	1	0	0	0	0	0	0	0	2
N_7	0	0	0	1	0	0	0	1	0	0	1	0	3
C_8	0	0	0	0	0	0	1	0	1	0	0	0	2
C_9	0	0	0	0	0	0	0	1	0	1	0	0	2
C_{10}	0	0	0	0	0	0	0	0	1	0	1	0	2
C_{11}	0	0	0	0	0	0	1	0	0	1	0	1	3
O_{12}	0	0	0	0	0	0	0	0	0	0	1	0	1

Table 2
Correlation weights of different vertices (chemical element together with the vertex degree) calculated by the Monte Carlo method for split 1, 2, and 3

V_k	$CW(V_k)$	Prevalence in training set	Prevalence in training set	Defect
Split 1, Eq.3				
C...1...	-0.15309	6	4	0.0071
C...2...	0.09204	13	6	0.0094
C...3...	1.92499	11	6	0.0021
Cl..1...	0.0	1	0	0.0000
N...1...	0.0	2	1	0.0000
N...3...	5.52455	7	3	0.0125
O...1...	0.57923	13	8	0.0034
O...2...	0.0	2	1	0.0000
S...2...	0.0	0	1	0.0000
Split 2, Eq. 4				
C...1...	-0.30200	6	5	0.0179
C...2...	0.12657	14	5	0.0197
C...3...	1.96087	11	6	0.0021
Cl..1...	0.86431	1	0	1.0000
N...1...	1.90949	2	2	0.0268
N...3...	5.40373	6	4	0.0071
O...1...	0.12862	13	7	0.0027
O...2...	0.62649	3	0	1.0000
Split 3, Eq. 5				
C...1...	-0.29507	6	6	0.0268
C...2...	0.11067	14	5	0.0197
C...3...	2.10422	11	8	0.0113
Cl..1...	0.0	0	1	0.0000
N...1...	1.13484	2	0	1.0000
N...3...	5.39626	6	4	0.0071
O...1...	0.58729	13	8	0.0034
O...2...	0.12783	3	0	1.0000

Table 3. The statistical characteristics of models for dispersibility of SWNTs in the organic solvents

	Training set (n=14)			**Calibration set (n=8)**			**Validation set (n=7)**		
Split	r^2	Q^2	F	r^2	$\overline{r_m^2}$	Δr_m^2	r^2	$\overline{r_m^2}$	Δr_m^2
1	0.605	0.420	18	0.885	0.83	0.04	0.900	0.81	0.09
2	0.611	0.436	19	0.888	0.67	0.15	0.953	0.88	0.05
3	0.607	0.440	19	0.931	0.90	0.00	0.912	0.59	0.19

Table 3. The experimental and calculated

ID*	SMILES	$Log_{10}C_{max}$, experiment	$Log_{10}C_{max}$, calculated	∑Defect	Domain of Applicability
	Split 1 $2\times\overline{\sum defect}$ =0.1320				
M02	O=C1N(C)CCCN1C	-0.1870	-0.1872	0.0730	YES
M05	CN1CCCC1=O	-0.9360	-1.3023	0.0533	YES
M09	N1(C(CCC1)=O)C=C	-1.0760	-1.2832	0.0627	YES
M14	O=CN(C)C	-1.6380	-1.7718	0.0396	YES
M16	CCC#N	-1.8240	-2.9880	0.0259	YES
M17	C=CC(=O)O	-1.8600	-2.3670	0.0254	YES
M20	C1CCC(=O)C1	-1.8890	-2.3982	0.0431	YES
	Split 2 $2\times\overline{\sum defect}$ =0.7097				
M02	O=C1N(C)CCCN1C	-0.1870	-0.1957	0.1140	YES
M05	CN1CCCC1=O	-0.9360	-1.4100	0.0890	YES
M09	N1(C(CCC1)=O)C=C	-1.0760	-1.3798	0.1087	YES
M14	O=CN(C)C	-1.6380	-2.0088	0.0653	YES
M17	C=CC(=O)O	-1.8600	-2.7257	0.0451	YES
M18	OCCSCCO	-1.8670	-3.0302	0.0843	YES
M20	C1CCC(=O)C1	-1.8890	-2.5941	0.0837	YES
	Split 3 $2\times\overline{\sum defect}$ =0.8251				
M06	O=C1CCCN1CCC#N	-0.9390	-0.9675	1.1402	No
M09	N1(C(CCC1)=O)C=C	-1.0760	-1.3250	0.1276	YES
M12	O=CN1CCCCC1	-1.4090	-1.6687	0.1290	YES
M14	O=CN(C)C	-1.6380	-1.9162	0.0839	YES
M16	CCC#N	-1.8240	-2.8780	1.0663	No
M17	C=CC(=O)O	-1.8600	-2.4378	0.0646	YES
M18	OCCSCCO	-1.8670	-2.7576	0.0858	YES

*) ID taken in Ref. [5]

Table 5.Correlation weights of different kinds of the vertex degrees obtained in three runs of the Monte Carlo calculations.

V_k	Run 1	Run 2	Run 3	Effect	Prevalence in Training set	Prevalence in Calibration set	Defect
Split 1							
C...2...	0.08953	0.09126	0.08627	increase	13	6	0.0094
O...1...	0.48932	0.53427	0.55085	increase	13	8	0.0034
C...3...	1.84634	1.79047	1.83472	increase	11	6	0.0021
N...3...	5.40255	5.42669	5.35392	increase	7	3	0.0125
C...1...	-0.20080	-0.20290	-0.19975	decrease	6	4	0.0071
N...1...	0.0	0.0	0.0	N/A*	2	1	0.0000
O...2...	0.0	0.0	0.0	N/A	2	1	0.0000
Cl..1...	0.0	0.0	0.0	N/A	1	0	0.0000
S...2...	0.0	0.0	0.0	N/A	0	1	0.0000
Split 2							
C...2...	0.12572	0.12594	0.14700	increase	14	5	0.0197
O...1...	0.16092	0.05264	0.73299	increase	13	7	0.0027
C...3...	1.90844	1.96168	2.37844	increase	11	6	0.0021
N...3...	5.27071	5.42916	6.00261	increase	6	4	0.0071
O...2...	0.61590	0.72085	0.42493	increase	3	0	1.0000
N...1...	1.77083	1.97330	2.00166	increase	2	2	0.0268
Cl..1...	0.92213	0.91521	0.78688	increase	1	0	1.0000
C...1...	-0.30019	-0.30069	-0.00482	decrease	6	5	0.0179
Split 2							
C...2...	0.11570	0.10125	0.12910	increase	14	5	0.0197
O...1...	1.21165	0.65059	1.10164	increase	13	8	0.0034
C...3...	2.62358	2.03992	2.52190	increase	11	8	0.0113
N...3...	5.99585	5.34519	5.99873	increase	6	4	0.0071
N...1...	1.24679	1.06117	1.38648	increase	2	0	1.0000
C...1...	-0.00109	-0.30296	0.00496	N/A	6	6	0.0268
O...2...	-0.12445	0.17736	-0.06459	N/A	3	0	1.0000
Cl..1...	0.0	0.0	0.0	N/A	0	1	0.0000

*) N/A = classification is not available

Author Contributions

A.P.T. had prepared the group of the random splits of available organic solvents into the training, calibration, and validation sets; had taken part in the carrying out the Monte Carlo experiments; and the discussion of the final text of the manuscript. A.A.T. had prepared the preliminary strategy of selection of group of versions for the Monte Carlo method and had prepared the preliminary version of the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References and Notes

1. Backes, C.; Hauke, F.; Schmidt, C.D.; Hirsch, A. Fractioning HiPco and CoMoCAT SWCNTs via density gradient ultracentrifugation by the aid of a novel perylene bisimide derivative surfactant *Chem. Commun.* **2009**, *19*, 2643–2645.
2. Choi, W.; Ohtani, S.; Oyaizu, K.; Nishide, H.; Geckeler, K.E. Radical polymer-wrapped SWNTs at a molecular level: High-rate redox mediation through a percolation network for a transparent charge-storage material. *Adv. Mater.* **2011**, *23*, 4440-4443.
3. Cheung, W.; Pontoriero, F.; Taratula, O.; Chen, A.M.; He, H. DNA and carbon nanotubes as medicine *Adv. Drug Deliv. Rev.* **2010**, *62*, 633-649.
4. Bergin, S.D.; Sun, Z.; Streich, P.; Hamilton, J.; Coleman, J.N. New solvents for nanotubes: Approaching the dispersibility of surfactants. *J. Phys. Chem.* C **2010**, *114*, 231-237.
5. Rofouei, M.K.; Salahinejad, M.; Ghasemi, J.B. An alignment independent 3D-QSAR modeling of dispersibility of single-walled carbon nanotubes in different organic solvents. *Fuller. Nanotube. Car. N.* **2014**, *22*, 605-617.
6. Yilmaz, H.; Rasulev, B.; Leszczynski, J. Modeling the dispersibility of single walled carbon nanotubes in organic solvents by quantitative structure-activity relationship approach. *Nanomaterials* **2015**, *5*, 778-791.
7. Gonzáles-Díaz, H.; Gia, O.; Uriarte, E.; Hernádez, I.; Ramos, R.; Chaviano, M.; Seijo, S.; Castillo, J.A.; Morales, L.; Santana, L.; Akpaloo, D.; Molina, E.; Cruz, M.; Torres, L.A.; Cabrera, M.A. Markovian chemicals "in silico" design (MARCH-INSIDE), a promising approach for computer-aided molecular design I: Discovery of anticancer compounds. *J. Mol. Model.* **2003**, *9*, 395-407.
8. Speck-Planche, A.; Scotti, M.T.; de Paulo-Emerenciano, V. Current pharmaceutical design of antituberculosis drugs: Future perspectives. *Curr. Pharm. Design* **2010**, *16*, 2656-2665.
9. Toropov, A.A.; Toropova, A.P.; Veselinović, A.M.; Veselinović, J.B.; Nesmerak, K.; Raska, I.; Duchowicz, P.R.; Castro, E.A.; Kudyshkin, V.O.; Leszczynska, D.; Leszczynski, J. The Monte Carlo method based on eclectic data as an efficient tool for predictions of endpoints for nanomaterials-two examples of application. *Comb. Chem. High. T. Scr.* **2015**, *18*, 376-386.
10. Worachartcheewan, A.; Prachayasittikul, V.; Toropova, A.P.; Toropov, A.A.; Nantasenamat, C. Large-scale structure-activity relationship study of hepatitis C virus NS5B polymerase inhibition using SMILES-based descriptors. *Mol. Divers.* **2015**, *19*, 955-964.
11. Toropova, A.P.; Toropov, A.A. CORAL software: Prediction of carcinogenicity of drugs by means of the Monte Carlo method. *Eur. J. Pharm. Sci.* **2014**, *52*, 21-25.

12. Toropova, M.A.; Toropov, A.A.; Raška, I.; Rašková, M. Searching therapeutic agents for treatment of Alzheimer disease using the Monte Carlo method. *Comput. Biol. Med*. **2015**, *64*, 148-154.
13. Toropov, A.A.; Toropova, A.P.; Gutman, I. Comparison of QSPR models based on hydrogen-filled graphs and on graphs of atomic orbitals. *Croat. Chem. Acta* **2005**, *78*, 503-509.
14. Gutman, I.; Toropov, A.A.; Toropova, A.P. The graph of atomic orbitals and its basic properties. 1. Wiener index. *MATCH Commun. Math. Comput. Chem*. **2005**, *53*, 215-224.
15. Gutman, I.; Furtula, B.; Toropov, A.A.; Toropova, A.P. The graph of atomic orbitals and its basic properties. 2. Zagreb indices. *MATCH Commun. Math. Comput. Chem*. **2005**, *53*, 225-230.
16. Toropov, A.A.; Toropova, A.P. Quasi-QSAR for mutagenic potential of multi-walled carbon-nanotubes. *Chemosphere* **2015**, *124*, 40-46.