Journal of Mathematical Chemistry Vol. 37, No. 4, May 2005 (© 2005)
DOI: 10.1007/s10910-004-1110-3

QSPR modeling of the enthalpy of formation based on Partial Order Ranking

Eduardo A. Castro, Francisco M. Fernández and Pablo R. Duchowicz*

INIFTA, División Química Teórica, Departamento de Química, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Diag. 113 y 64, Suc. 4, C.C. 16, (1900) La Plata, Argentina
E-mail: duchow@inifta.unlp.edu.ar

Received 13 September 2004; revised 28 September 2004

A new predictive method based on Partial Order Ranking is introduced in the realm of the QSPR–QSAR Theory using a single descriptor as variable and which is simple enough to perform calculations by hand. Comparisons are made with a model relying on the Least Squares Method subjected to the modeling of the Enthalpy of Formation from Elements exhibed by a set of 51 hydrocarbon molecules by means of a flexible type descriptor. The results achieved with the proposed method are quite satisfactory and its future aplicability seems to be very promising.

KEY WORDS: QSPR–QSAR theory, least squares method, partial order ranking, flexible descriptor, enthalpy of formation from elements

1. Introduction

The Quantitative Structure Property–Activity Relationships (QSPR–QSAR) have emerged in the last decades as a useful tool for predicting physicochemical, biological and pharmacological properties of molecules, specially in those cases where there are no available experimental data corresponding to such properties [1,2]. Since Hansch and Fujita made the pioneer studies in 1964 [3], these have been succesfully applied in the estimation of different properties and activities and so the development of the theory is encouraged.

The different formulations of the QSPR–QSAR theory suggest mathematical models quantifying an hypothetical relationship between the molecular structure and the property p shown by a compound,

$$\mathrm{p} = \mathrm{function}\{d\}, \tag{1}$$

where $\{d\}$ denote a set of molecular descriptors describing the molecular structure/substructure. The $\{d\}$ variables often represent experimentally

* Corresponding author.

determined physicochemical properties [2] or theoretically derived quantities, for example, from the Chemical Graph Theory [4].

In order to obtain the model indicated with (1) the first issue to be established is a structural function relating the property under study with the set of descriptors, which is usually an unknown relationship depending on the property, the descriptors and the family of molecules being studied. It can represent a linear or a non-linear dependence. The simplest way to elucitate a mathematical function is to draw the dependent variable in the z-axis of a cartesian coordinate reference system and the descriptors in the other axes. Obviously, this operation can be done for no more than two descriptors.

A second point to cover is the choice of a statistical method that enables the calculation of all the function parameters and finally the prediction of the property. Different methods have been used as data reduction techniques, such as: Multiple Regression Analysis (MRA) [5], Principal Component Analysis (PCA) [6], Partial Least Squares (PLS) [7], or Artificial Neural Networks (ANN) [8]. The MRA is the commonest used technique for many years in the field of QSPR–QSAR Theory owing to the simplicity of its equations, representing the other advanced methodologies introduced in the last 30 years which developed upon MRA. Although the ANN is often a heavily parametrized method it has the advantage that it does not need a structural function to be specified, since the algorithms try to guess it.

Another important point to bear in mind is the selection of the molecular descriptors. The complexity of the molecular structure has not allowed the design of a novel descriptor englobing the whole structural information into a single variable [9]. Nowadays there are more than 1000 molecular descriptors available in the standard literature [10,11] and the researcher has to deal with the problem of selecting the best subset from an initial pool of descriptors [12–14]. There is nothing definitively stated in this regard, and the best way of doing that is to choose those descriptors generating the best predictions.

As soon as all these topics are fulfilled a calibration set of compounds can be modeled. The criterium to define the quality of a model can be the functional standard deviation of estimate S or the Fischer test parameter. The model must assure that it has predictive ability for a set of compounds not included in the calibration set and comprising the validation set. When there are no reasons to doubt about it, the leave-one-out cross validation technique gives the real predictive power of the model [15].

An alternative atractive methodology for designing a QSPR–QSAR study is the Partial Order Ranking method, a technique which does not rely on a structural function for the model nor a statistical method for the predictions like the already mentioned above. Its theory has been described in several papers [16–18] and is based on elementary methods of Discrete Mathematics, appearing from a mathematical point of view extremely simple compared to the more demanding statistical methods such as MRA or PCA/PLS.

The purpose of this paper is to present a new predictive method based on Partial Order Ranking for a single descriptor. Since up to our knowledge the literature does not register the use of just only one variable for employing Partial Order Ranking, the results obtained in this paper are in principle the first in the subject. The predictions given by this new method are compared with the results given in ref. 19. The article is organized as follows: next section describes the Partial Order Ranking method. Then a discussion of the results obtained is presented. Finally, we summarize the main conclusions of this work and suggest some possible future extensions of the new technique.

2. Method

The methodology of Partial Order including a single descriptor has an extremely simple principle: if a molecule j with a given property p_j is characterized with a descriptor d_j, then two molecules A and B can be compared if and only if their descriptors can be compared. That is to say,

$$p_B \leqslant p_A \leftrightarrow d_B \leqslant d_A. \tag{2}$$

When the rule (2) is true then it is said that compound A is ranked higher than compound B. If (2) is false, then both A and B are incomparable. Note that (2) *a priori* includes "$\leqslant$" as the only structural function.

First of all consider a calibration set a with N compounds. If we apply (2) to this set then it will generate two different subsets a_1 and a_2: in a_1 all the molecules will satisfy (2) and the second subset a_2 will contain those compounds which dont follow the rule. However, if we apply again (2) to a_2 we will generate two new different subsets a_3 and a_4, with fewer elements each one, where compounds in a_3 are ordered and compounds in a_4 do not obey rule (2). Proceeding in this way repeatedly, we continue iterating until the number of compounds in the second subset is zero. This condition is achieved only if the selected descriptor can describe the whole calibration set. Otherwise, the second subset of the last iteration will not be empty. After all this procedure is done, we will have the following k ordered subsets $a_h / h = 1, \ldots, k$, and where k is dependent on the property p_j under consideration and the descriptor d_j employed.

In order to predict the property p_i of a given compound i with descriptor value d_i from the calibration set using the k ordered subsets, we can use simple interpolation formulae. First, we have to locate the subset a_x that contains a compound j (with i excluded from a_x) that satisfies the next condition

$$\text{absolute}(d_j - d_i) = \text{minimum}. \tag{3}$$

Once located the subset a_x and the molecule j, then the following situation will appear in a_x,

$$\begin{array}{ll} p_j & d_j \\ & d_i \\ p_{(j+1)} & d_{(j+1)}, \end{array}$$

where we have the ranking $j \leqslant i \leqslant j+1$.

The linear interpolation formulae can be deduced as

$$\begin{gathered} p_j = a * d_j, \\ p_{(j+1)} = a * d_{(j+1)}, \\ p_i = a * d_i, \\ p_{(j+1)} - p_j/(d_{(j+1)} - d_j) = a, \\ p_i(\text{pred}) = p(j+1) - p_j/(d_{(j+1)} - d_j) * d_i \end{gathered}$$

with p_i(pred) denoting the predicted value of p_i. For the special case where $p_j = p_{\min}$ or $p_j = p_{\max}$, where $p_{\min}$ and $p_{\max}$ are the minimum and maximum values of p_j in a_x, respectively, we can generalize the previous equations,

$$\begin{gathered} p_j = a * d_j \\ p_j/d_j = a, \\ p_i(\text{pred}) = (p_j/d_j) * d_i, \end{gathered}$$

If we have a validation set we proceed in a similar way as indicated above: first, localizating the minimum difference between the descriptor d_i of the validation set and a d_j from a subset a_x according to the condition (3), and then applying the linear interpolation formulae.

It could be shown that the lower the value of k the better are the estimations of the proposed methodology. This can be concluded from the fact that condition (3) is not sufficient to lead to the best predictions for a descriptor d_i in a_x when k is greater. Another point is the length of the interval $d_{(j+1)} - d_j$: the lower the length of the interval, the better the predictions. This is in consecuence of the linear interpolation formulae: a secant line is approximating a tangent line in a property versus single descriptor-graph.

3. Results and discussion

In ref. 19 of the work that we based our calculations the enthalpies of formation from elements ΔH_f° (kcal/mol) of 51 hydrocarbons where partitioned into a calibration and a validation set, each composed of 36 and 15 molecules, respectively. However, owing to the results of the work of Hawkins et al. [15], we decided to use the complete set of 51 molecules for calibration, since we

do not have any reason to doubt of the predictive hability of the leave-one-out technique.

If we calibrate the model using linear least squares and a flexible descriptor such as the correlation weighting of local invariants of atomic orbital molecular graphs that was used in the reference, the best model found resulted for probe number 2 from the three probes reported.

Model 1.

$$\Delta H_f^\circ = -4.264 + 1.394 * \text{Oxc (probe 2)},$$
$$R_{cal} = 0.9989;\ S_{cal} = 1.8114;\ F = 2.2102 \times 10^4;\ \text{absqdev} = 3.1526,$$
$$R_{lou} = 0.9988;\ S_{lou} = 1.8533.$$

Here R_{cal}, S_{cal} and F are the correlation coefficient, standard deviation and Fisher test parameter of the calibration set, respectively; absqdev is the mean absolute quadratic deviation of the model; R_{lou} and S_{lou} stand for the correlation coefficient and standard deviation of the leave-one-out cross validation method.

Applying partial order we get the predictions for the 51 molecules and $k = 2$, that is, 2 subsets totally ordered listed in table 1. In order to compare these predictions with the previous result of least squares, we did a correlation between the experimental values of the property and the predicted values using partial order. We got that the best model results for probe 1,

Model 2.

$$\Delta H_f^\circ = -0.254 + 1.012 * \text{Oxc (probe 1)},$$
$$R_{cal} = 0.9974;\ S_{cal} = 2.7695;\ F = 9426.7;\ \text{absqdev} = 7.3698.$$

In table 2 we display the predicted values for the heats of formation with models 1 and 2. Note that descriptors Oxc (probe 1) and Oxc (probe 2) are not the same.

When judging the quality of the model with the standard error of the leave-one-out method, it can be concluded that the new method produced worse predictions. This result can be explained arguing that the descriptor Oxc may not be optimum for partial order. If the descriptor performs well in a regression, it does not mean that it will also work well with partial order. This is why the best model chooses two different Oxc in model 1 (probe 2) and in model 2 (probe 1). The opposite situation is also valid: if a descriptor orders itself just like the property does, then it does not mean a good correlation between the descriptor and the property is expected. This last conclusion can be demostrated with a simple numerical experiment: if we order the set of 51 enthalpies and create an arbitrary mathematical descriptor D_0 by taking the square root of j, $j = 1, \ldots, 51$, then we will find a better statistic with partial order instead of that provided when using that descriptor in a linear regression model.

Table 1
List of ordered subsets generated according to condition (2) for 51 hydrocarbon molecules.

Subset a_1			Subset a_2		
ID	ΔH_f° (kcal/mol)	Oxc(probe1)	ID	ΔH_f° (kcal/mol)	Oxc (probe1)
20	−40.140	−49.860	27	−39.940	−50.970
40	−36.600	−46.350	38	−30.330	−41.730
13	−32.240	−40.940	17	−18.260	−23.110
7	−24.930	−31.000	1	−17.790	−23.080
23	−23.670	−30.240	51	4.310	−4.112
37	−20.040	−18.870	34	4.560	3.862
43	−12.400	−14.930	6	4.790	4.431
12	−4.270	−6.246	42	29.900	28.940
11	−1.770	−3.784	45	36.000	38.220
10	0.070	−1.514	19	37.450	43.760
50	4.140	5.278	44	43.560	47.530
16	8.440	8.770	41	51.900	59.550
36	11.950	13.050	29	55.200	62.550
3	12.560	13.600	18	66.220	78.890
39	18.290	17.100	33	69.200	81.140
15	25.270	29.060	25	79.700	95.200
8	26.010	29.830			
14	32.120	37.460			
9	34.690	40.590			
35	35.400	41.980			
32	37.230	43.770			
22	37.700	44.250			
49	43.300	48.300			
24	44.250	52.200			
4	44.410	52.280			
5	45.310	53.510			
46	49.700	59.660			
31	54.000	64.380			
2	54.550	64.700			
28	59.180	70.250			
26	62.500	74.280			
47	66.000	81100			
30	68.100	83.990			
48	78.400	99.250			
21	148.690	178.900			

Linear Least Squares

$$\Delta H_f^\circ = -25.349 + 199.767 * D_0,$$
$$R_{cal} = 0.8316;\ S_{cal} = 21.3870;\ F = 109.905;\ \text{absqdev} = 439.470,$$
$$R_{lou} = 0.8081;\ S_{lou} = 22.5032.$$

Table 2
Experimental and predicted values of Entalpies of Formation from Elements (kcal/mol) for 51 hydrocarbons with models 1 and 2.

ID	Oxc (probe 1)	Oxc (probe 2)	ΔH_f° exp	ΔH_f° model 2	ΔH_f° model 1
1	−23.080	−9.316	−17.79	−18.230	−17.250
2	64.700	42.070	54.55	54.280	54.370
3	13.600	12.330	12.56	12.800	12.920
4	52.280	34.800	44.41	44.310	44.240
5	53.510	35.480	45.31	45.280	45.190
6	4.431	6.712	4.79	5.134	5.089
7	−31.000	−14.390	−24.93	−24.280	−24.320
8	29.830	21.470	26.01	25.890	25.660
9	40.590	27.860	34.69	34.380	34.560
10	−1.514	3.286	0.07	−0.290	0.314
11	−3.784	1.990	−1.77	4320	−1.492
12	−6.246	0.422	−4.27	1.824	−3.678
13	−40.940	−19.740	−32.24	−29.810	−31.780
14	37.460	26.090	32.12	35.490	32.100
15	29.060	21.160	25.27	29.980	25.230
16	8.770	9.176	8.44	7.646	8.524
17	−23.110	−9.530	−18.26	−17.800	−17.550
18	78.890	50.420	66.22	64.860	66.010
19	43.760	29.650	37.45	37.220	37.060
20	−49.860	−25.260	−40.14	−38.770	−39.470
21	178.900	108.700	148.60	141.300	147.200
22	44.250	30.210	37.70	37.870	37.840
23	−30.240	−13.560	−23.67	−24.620	−23.160
24	52.200	34.620	44.25	44.380	43.990
25	95.200	59.840	79.70	75.660	79.140
26	74.280	47.680	62.50	61.710	62.190
27	−50.970	−25.210	−39.94	−41.040	−39.400
28	70.250	45.250	59.18	59.150	58.810
29	62.550	40.320	55.20	52.330	51.930
30	83.990	53.790	68.10	71.320	70.710
31	64.380	41.960	54.00	54.240	54.220
32	43.770	29.680	37.23	37.460	37.100
33	81.140	51.420	69.20	66.020	67.410
34	3.862	6.500	4.56	4.758	4.794
35	41.980	28.580	35.40	35.800	35.570
36	13.050	11.770	11.95	12.090	12.140
37	−18.870	−10.470	−20.04	−15.290	−18.860
38	−41.730	−21.400	−30.33	−32.870	−34.090
39	17.100	17.810	18.29	15.440	20.560
40	−46.350	−23.300	−36.60	−37.030	−36.740
41	59.550	39.020	51.90	49.620	50.120
42	28.940	21.390	29.90	25.190	25.550
43	−14.930	−4.754	−12.40	−15.120	−10.890
44	47.530	32.860	43.56	42.240	41.530
45	38.220	26.820	36.00	32.740	33.110

Table 2 *(Continued)*

ID	Oxc (probe 1)	Oxc (probe 2)	ΔH_f° exp	ΔH_f° model 2	ΔH_f° model 1
46	59.660	40.280	49.70	52.010	51880
47	81.100	53.750	66.00	69.140	70.650
48	99.250	62.380	78.40	83.090	82.680
49	48.300	33.590	43.30	44.090	42.550
50	5.278	8.798	4.14	5.657	7.997
51	−4.112	3.242	4.31	−2.103	0.253

Partial Order Ranking

$$\Delta H_f^\circ = -0.306 + 1.029 * \mathrm{D0},$$
$$R_{cal} = 0.9878;\ S_{cal} = 6.0062;\ F = 1965.8;\ \mathrm{absqdev} = 34.660.$$

The nature of the new method suggested reveals two important features:

1. Since (2) includes "$\leqslant$" as the only structural function, this structural function will always be true whenever the descriptor used obey this rule. This mean that, in comparison to classical least squares method, one does not have to look for this structural function but for the descriptors that behave correctly.
2. The proposed method does not need to be validated with the leave-one-out technique. That is so because when the property of compound i is being predicted with the interpolation formulae, it is really being practised the leave-one-out technique, as compound i is being left out of the entire calibration set and being predicted with its immediate neighbors, j and $j + 1$. A leave-one-out like this should function better than the ordinary leave-one-out obtained with the least squares method, since it also does not depend on the structural function. In other words, the R_{cal} and S_{cal} of Partial Order "correspond to R_{lou} and S_{lou} of Least Squares" whenever the structural function being used in the model is valid for least squares.

4. Conclusions

Partial Order Ranking can be considered as a parameter-free method, and so neither assumptions about linearity nor assumptions about distribution properties are made. A structural function is difficult to elucitate in cases where the property has a complicated dependence on the molecular features. Here is when the utility of Partial Order appears: if the descriptor orders itself just like the property values do, even when treating with difficult properties the method would perform well. The only difficult part in all this could appear when trying

to find molecular descriptors which order themselves accordingly to the property. There are different ways to surmount this problem and one of them is to optimize flexible type descriptors taking into account their ranking with the property. This kind of experiments are being carried out in our laboratories and the corresponding results will be presented elsewhere in the near future.

References

[1] W.A. Sexton, *Chemical Constitution and Biological Activity*, (D. Van Nostrand, New York, 1950)
[2] C. Hansch, Acc. Chem. Res. 2 (1969) 232.
[3] C. Hansch and T. Fujita, J. Am. Chem. Soc. 86 (1964) 1616.
[4] R.B. King, (ed.), Chemical applications of topology and graph theory, *Studies in Physical and Theoretical Chemistry*, Vol. 28 (Elsevier, Amsterdam, 1983 E. R.).
[5] Malinowski, *Factor Analysis in Chemistry* (Wiley, New York, 1991).
[6] H. Hotelling, J. Educ. Psychol. 24 (1933) 417.
[7] S. Wold, M. Sjostrom and L. Eriksson, in: *Encyclopedia of Computational Chemistry*, eds. R. von Schleyer, N.L. Allinger, T. Clark, J. Gasteiger, P.A. Kollman, H.F. III Schaefer, P.R. Schreiner, Vol. 3, (Wiley, Chichester, England, 2006, 1998).
[8] J. Zupan, in: *Encyclopedia of Computational Chemistry*, Vol. 33, eds. R. von Schleyer, N.L. Allinger, T. Clark, J. Gasteiger, P.A. Kollman, H.F. Schaefer, III, P.R. Schreiner, (Wiley, Chichester, England, 2006, 1998).
[9] L.B. Kier and L.H. Hall, *Molecular Connectivity in Chemistry and Drug Research* (Academic Press, New York, 1976).
[10] R. Todeschini, *Handbook of Molecular Descriptors* (Wiley-VCH, Berlin, 2000).
[11] S.C. Basak, D.K. Harris and V.R. Magnuson, POLLY (version 2.3), Copyright of the University of Minnesota, 2001.
[12] D. Rogers and A.J. Hopfinger, J. Chem. Inf. Comput. Sci. 34 (1994) 123.
[13] A. Hoskuldsson, Chemom. Intell. Lab. Syst. 32 (1996) 37.
[14] R. Hoffmann, V.I. Minkin and B.K. Carpenter, Bull. Soc. Chim. Fr. 133 (1996) 117.
[15] D.M. Hawkins, S.C. Basak and D. Mills, J. Chem. Inf. Comput. Sci. 43 (2003) 579.
[16] L. Carlsen, P.B. Sorensen and M. Thomsen, Chemosphere 43 (2001) 295.
[17] R. Bruggemann, S. Pudenz, L. Carlsen, P.B. Sorensen, M. Thomsen and R.K. Mishra, SAR-QSAR Environ. Res. 11 (2001) 473.
[18] L. Carlsen, P.B. Sorensen, M. Thomsen and R. Bruggemann, SAR QSAR Environ. Res. 13 (2002) 153.
[19] A. Mercader, E.A. Castro and A.A. Toropov, Chem. Phys. Lett. 330 (2000) 612.