Eur. J. Med. Chem. 33 (1998) 659–664

Autocorrelation modeling of lipophilicity with a back-propagation neural network

James Devillers*, Daniel Domine, Cécile Guillon

CTIS, 3 Chemin de la Gravière, 69140 Rillieux La Pape, France

(Received 17 November 1997; accepted 9 March 1998)

Abstract – From a training set of 7200 chemicals a back-propagation neural network (BNN) model was developed for estimating the *n*-octanol/water partition coefficient of organic molecules. Chemicals were described by means of a modified autocorrelation method. The advantages of the autocorrelation method were emphasized through the analysis of the simulation performances of the model and from a comparative study involving another BNN model [Quant. Struct. Act. Relat. 16 (1997) 224-230] using a large number of variables (atoms and bonds) derived from connection matrices. © Elsevier, Paris

***n*-octanol/water partition coefficient / autocorrelation method / back-propagation neural network / AUTOLOGP™ (Version 4.0)**

1. Introduction

The *n*-octanol/water partition coefficient (log *P*) is the molecular descriptor of choice in QSAR for simulating lipophilicity of organic molecules. However, even if experimental log *P* values are available for a huge number of chemicals, it is impossible to envision the measure of this parameter for all the existing chemicals and those designed in lead optimizations. Under these conditions, numerous calculation procedures have been proposed in the literature. The most common approaches for estimating log *P* are classified as 'fragmental methods' in which a molecule is divided into atoms and/or functional groups having specific numerical values encoding lipophilicity [e.g., 1, 2]. The summation of these values (generally with correction factors) yields the log *P* estimate. The most serious limitations of the fragmental approaches are the difficulty of defining valuable group contributions, the lack of physicochemical basis for some fragments, and the justification of correction factors which often seem to appear in the calculations just at the right time! To overcome these problems, attempts have been made for deriving log *P* models directly from molecular properties [e.g., 3–5] rather than fragments. These methods obviate the necessity to define and justify fragments and allow computation of log *P* for different conformers of the same chemical [6]. However, the obtained models often have limited validity outside their chemical training sets [6] and they cannot be used for all the molecules. Topological indices [7] have been employed for facilitating the description of the molecules and therefore, for deriving log *P* models with a wider domain of application [8–11]. Unfortunately, most of these topological indices are redundant and difficult to interpret. The autocorrelation descriptors, introduced by Moreau and Broto [12–14] present the advantages of the classical topological indices but they do not suffer from their limitations. Indeed, they can be calculated for any organic molecule. In addition, they are weakly redundant and can be linked to physicochemical properties (e.g., lipophilicity, electronegativity). Under these conditions, attempts have been made for deriving log *P* models from these parameters [15–20]. In this paper, we employed different 2-D autocorrelation descriptors for the design of a new model allowing the prediction of log *P* values for structurally diverse organic molecules.

*Correspondence and reprints

2. Materials and methods

2.1. Experimental data

Measured log P values for the 7200 molecules constituting the training set and the 519 molecules of the external testing set were retrieved from original publications (i.e. journals, theses, reports) and unpublished results. The log P values of the training set ranged between – 3.7 to 9.95 with a mean of 2.13 and a standard deviation of 1.65. For the testing set, the smallest log P value was – 1.62 and the largest 10.2. The mean and standard deviation were 2.89 and 1.98, respectively. A validation set of 200 molecules was also designed for selecting the optimal configuration of the BNN. These 200 experimental log P data were also retrieved from original articles. The validation set was constituted of chemicals presenting a high structural diversity but not deviating too much from the 7200 chemical structures included in the training set.

2.2. Autocorrelation descriptors

In the autocorrelation method, a molecule is represented by means of a graph where the atoms are displayed by nodes and bonds are depicted by edges. In this graph, the distance between two nodes is defined as the smallest number of edges between them. Considering that a property of a molecule can be calculated from atomic contributions (AC_i) describing this molecule, the classical autocorrelation algorithm allows to compute all the products ($AC_i \times AC_i$; $AC_i \times AC_j$;...) corresponding to the different smallest internodal distances (i.e., 0 to n) in the molecular graph. The sum of these products for the same distance in the graph gives a component of the autocorrelation vector for the selected property. A detailed description of the autocorrelation algorithm with examples of calculation can be found in the literature [12–14, 21]. To optimize the descriptive power and the weak redundancy of the autocorrelation vectors, we defined values for specific functional groups and designed a new autocorrelation algorithm [16, 17]. It principally allows to obtain the first component of an autocorrelation vector by simply summing the positive and negative contributions attributed to the atoms and functional groups constituting the studied molecule.

Molecules constituting the three different sets were described by means of four different autocorrelation vectors. First, from the fragmental constants of Rekker and Mannhold [1], it has been possible to compute for each molecule, an autocorrelation vector H representing lipophilicity. Second, an autocorrelation vector MR, encoding molar refractivity, was designed from the fragmental constants of Hansch and Leo [22] or directly from the Lorentz-Lorenz equation (Eq. (1)):

$$MR = \frac{n^2 - 1}{n^2 + 2} \times \frac{MW}{d} \qquad (1)$$

where n was the refraction index, d the density, and MW the molecular weight of the molecule.

Last, the H-bonding acceptor ability and H-bonding donor ability of the molecules were described by means of Boolean contributions. The resulting autocorrelation vectors were termed HBA and HBD, respectively.

2.3. BNN analysis

Due to the high dimensionality of our training set, a three-layer back-propagation neural network (BNN) including a bias in the input layer and hidden layer was used as statistical engine [23, 24].

Data were scaled with the following equation:

$$x'_i = [a\ (x_i - x_{min}) / (x_{max} - x_{min})] + b \qquad (2)$$

where x'_i was the scaled value, x_i the original value, x_{min} and x_{max} the minimum and maximum values of each column, respectively. In order to have scaled data ranging between 0.05 and 0.95 the values of a and b equaled 0.9 and 0.05, respectively.

Numerous assays were performed to set the optimal architecture of the BNN. Thus, we have tried to optimize the number of input (i.e., autocorrelation components) and hidden neurons, the values of the learning rate (η) and momentum term (α), and the number of learning cycles necessary to obtain good generalization performances. The training exercises were monitored with the validation set of 200 structurally diverse molecules.

3. Results and discussion

Thirty-five autocorrelation descriptors were necessary to correctly describe the molecules and model log P. These input neurons were constituted of the 15 first components of the autocorrelation vector H encoding lipophilicity, the 15 first components of the autocorrelation vector MR representing molar refractivity, the four first components of the autocorrelation vector encoding the H-bonding acceptor ability (HBA) of the molecules, and the first component of the autocorrelation vector HBD. Thirty-two neurons on the hidden layer, a learning rate (η) of 0.5 and a momentum term (α) of 0.9 always allowed us to obtain good BNN generalizations within ~ 5500 cycles. In our study, different log P models presenting a high predictive power were obtained. Under

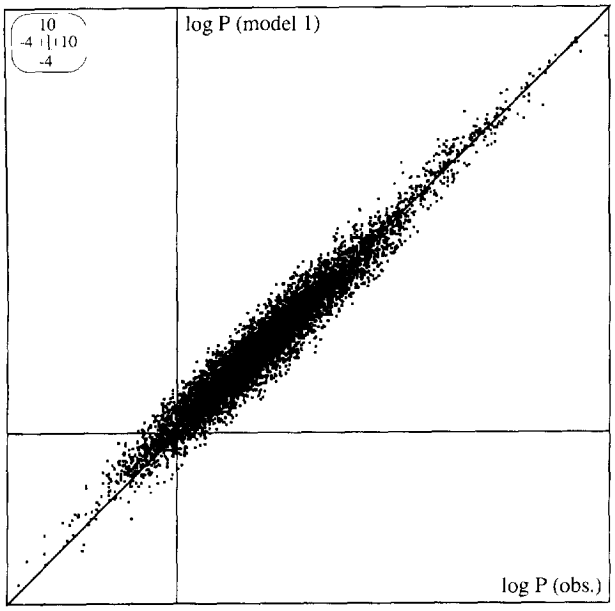

Figure 1. Calculated log P values (model 1) compared to the experimental (obs.) data used in the training set of model 1 (7200 molecules).

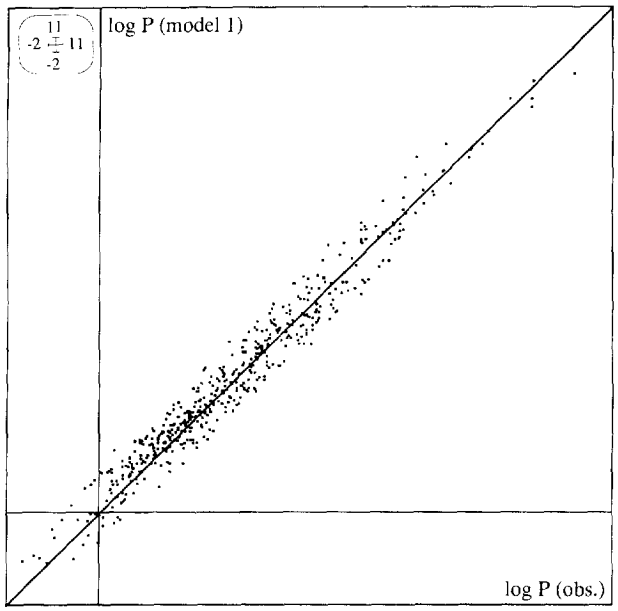

Figure 2. Calculated log P values (model 1) compared to the experimental (obs.) data used in the testing set of model 1 (519 molecules).

these conditions, an ensemble of networks [25] constituted of valuable configurations was designed by a trial and error procedure consisting in the evaluation of the best combination of configurations allowing to obtain reliable estimations of log P values of the testing set (i.e. average performance [25]). Thus, a composite network constituted of four configurations was selected as final model (root mean square error (RMS) = 0.37, $r = 0.97$). It allowed to obtain the best simulation results with the testing set (RMS = 0.39, $r = 0.98$). Plots of the experimental versus calculated log P values for the compounds constituting the training and testing sets are shown in *figures 1* and *2*, respectively. These figures reveal good agreement between the experimental and calculated log P values for the 7719 molecules.

A fundamental principle guiding the selection of a model is that enough information is provided for estimating its domain of application. Thus, our autocorrelation/BNN model allows the calculation of log P values of any organic molecule containing nitrogen, oxygen, halogen, phosphorus, and/or sulfur atoms. Due to the selected methodology, the model appears well suitable for simulating the lipophilicity of molecules presenting a high degree of structural diversity. In addition, a particular attention has been paid for encoding branching information of the molecules. However, it is important to note that the log P values of salts, amino acids, nucleosides, and nucleobases cannot be calculated because the model has not been designed for these compounds. Also, the log P values of some chemicals with formal charges (such as nitrogen oxides) cannot be estimated, with the exception of nitro compounds. Last, the log P values of ionisable chemicals (e.g., carboxylic acids) can be calculated although pH-dependencies should be noted.

In order to test the performances of our model, a comparative study was carried out with a recent published BNN model [26]. This model was designed from a training set of 268 molecules constituted of a maximum of ten non-hydrogen atoms and tested on 50 different compounds. The log P values were retrieved from the compilation of Hansch and coworkers [27]. Molecules were described by means of their connection matrices in which atoms are represented horizontally and vertically along the edges. The diagonal and non-diagonal entries are the atoms and bonds, respectively. A unique (i.e., canonical) numbering of the atoms in the molecule was obtained by applying the Moreau's method (explained in [26]). Atom and bond types (i.e., each entry in the connection table) were described by means of eight (for C, N, O, S, F, Cl, Br, and I atoms) and four (for single, double, triple, and aromatic/nitro types) indicator variables, respectively. Using the canonical numbering, 260 variables were therefore generated among which those containing only zero entries were removed, yielding 147 descriptors. A three-layer BNN was used. Prior to calculations, data were scaled to lie between -0.9 and 0.9. The best predictive performances of the neural network were obtained when using three neurons in the

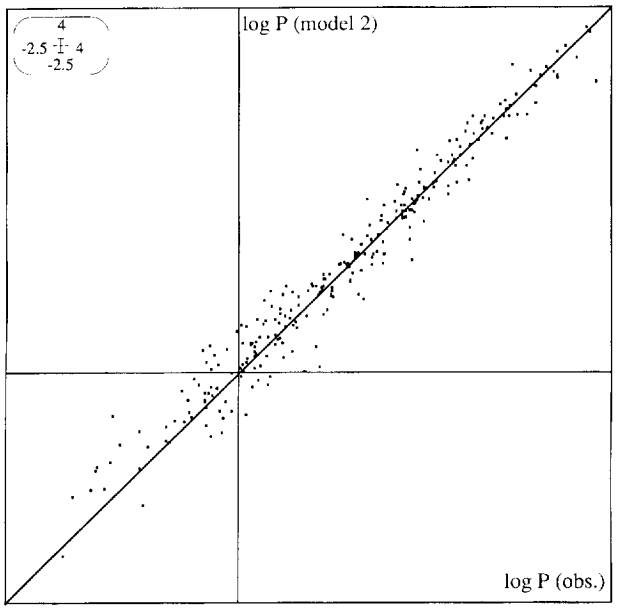

Figure 3. Calculated log P values (model 2) compared to the experimental (obs.) data used in the training set of model 2 (265 molecules).

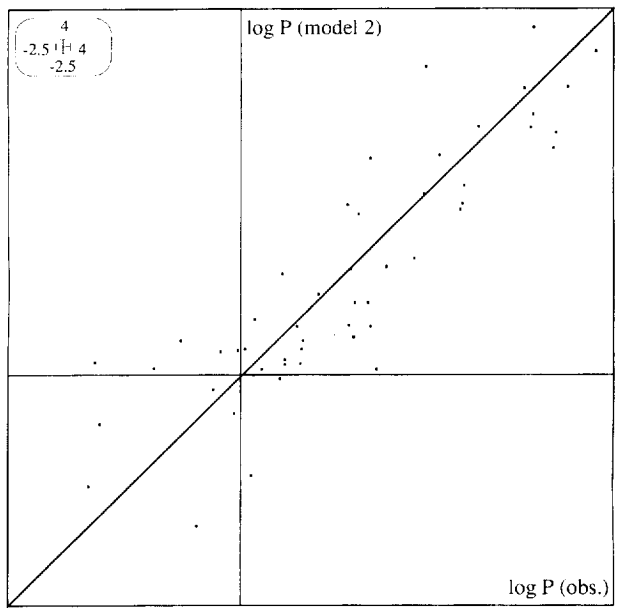

Figure 4. Calculated log P values (model 2) compared to the experimental (obs.) data used in the testing set of model 2 (50 molecules).

hidden layer. The learning rate (η) was initially set at 0.05 and reduced to 0.01, 0.005 or 0.002 if oscillations were observed. A momentum term (α) of 0.7 was always used. The number of learning cycles was not specified. For the training set, a RMS of 0.25, an $s = 0.25$, and an $r = 0.98$ were obtained. For the testing set, these parameters equaled 0.65, 0.66, and 0.88, respectively. Note that to obtain the same figures as Schaper and Samitier [26], it is necessary to round their calculated values to two decimals. *Figures 3* and *4* illustrate the performances of the Schaper and Samitier model for their training and testing sets, respectively.

The performances of the two models were assessed and compared on the basis of the training and testing sets of model 2 [26]. Indeed, for this model, it was impossible to calculate the log P values of new compounds due to the fact that Schaper and Samitier did not provide enough information (no information was given either regarding the commercial availability of their model). In addition, the log P values used in their study were employed in order to perform a fair comparison of the two models although some selected values are discussible.

For a first and rapid comparison, the observed versus calculated log P values were plotted for the training set (*figures 3* and *5*) and testing (*figures 4* and *6*) set of the Schaper and Samitier model. *Figures 3* and *5* illustrate the good performances of the two BNN models for the compounds belonging to the training set of model 2. However, *figures 4* and *6* show the superiority of our model for the compounds used as testing set when deriving the model of Schaper and Samitier [26]. This difference cannot be only explained by the fact that our model was derived from a larger training set than their model. Indeed, the obtained results stress a major problem in the development of the model of Schaper and Samitier. It is well known that BNNs can overfit the training data when the number of individuals is too low compared with the number of connections within the network and/or when the number of training cycles too large [24]. In such instances, the BNNs present a poor generalization ability. In the model of Schaper and Samitier, the ratio of the number of training individuals over the number of connections in the network equals 0.6 (note that this point has been stressed by Schaper and Samitier [26]) so that the BNN overfits the training data and cannot correctly predict the log P values of the test compounds.

Computation of the RMS value, the standard error of estimate (s), and the correlation coefficient (r) for the two models (*table I*) confirms the above conclusions. They show that the model of Schaper and Samitier (model 2) provides better statistics than our model (model 1) for the 265 compounds of their training set. As stressed above, this is due to the fact that the BNN was allowed to fit the data by using a large number of variables (inducing a large number of connections). These statistics nevertheless reveal that our model provides acceptable estimations of log P for this set of chemicals since the RMS and

Table I. Statistics for model 1 (this study) and model 2 [26].

Model	RMS	r	s
Training set of model 2 (265 chemicals) [a]			
Model 1	0.30	0.97	0.30
Model 2	0.25	0.98	0.25
Testing set of model 2 (50 chemicals)			
Model 1	0.37	0.96	0.38
Model 2	0.65	0.88	0.66

[a] The current version of AUTOLOGP™ (implementing model 1) is not able to calculate the log P value of methane, water, and formic acid included in the training set of model 2 (compounds 1, 22, and 56, respectively in [26]).

s values are below 0.4. For the 50 compounds belonging to the testing set, our model still provides RMS and s values below 0.4 but the model of Schaper and Samitier [26] fails to give accurate estimations.

4. Conclusion

Autocorrelation vectors represent valuable molecular descriptors since they can be computed for any arbitrary organic molecule and the calculations are made without error. In addition, these algorithmically derived parameters are weakly redundant and are easily interpretable.

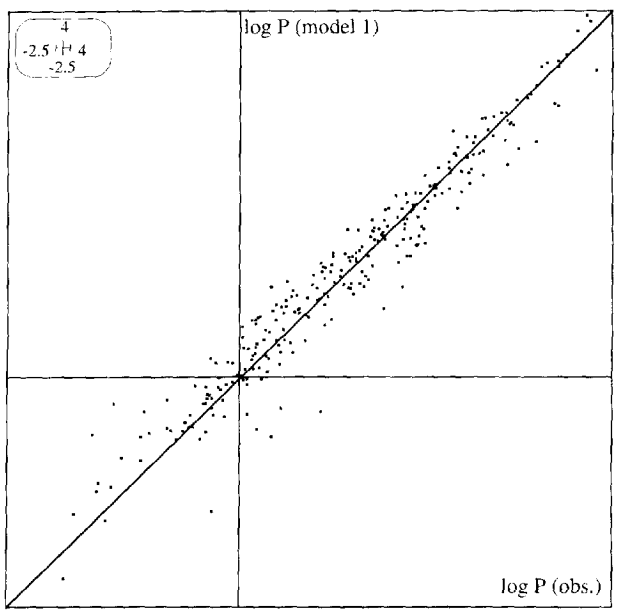

Figure 5. Calculated log P values (model 1) compared to the experimental (obs.) data used in the training set of model 2 (265 molecules).

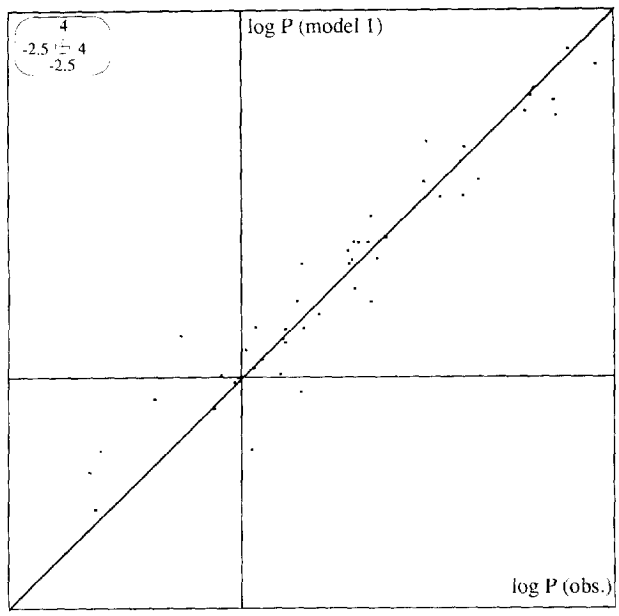

Figure 6. Calculated log P values (model 1) compared to the experimental (obs.) data used in the testing set of model 2 (50 molecules).

Indeed, the autocorrelation method allows to produce molecular descriptors with a very low degree of degeneracy. The use of atomic contributions encoding various physicochemical properties of the molecules facilitates the interpretation of the derived models. Under these conditions, it is not surprising that the autocorrelation method has clearly shown its efficiency in numerous QSAR and QSPR studies [e.g., 28–34]. The usefulness of the autocorrelation vectors introduced in a back-propagation neural network for modeling the n-octanol/water partition coefficient of the organic molecules has been shown in a previous work [20] and is clearly confirmed in the present study. Indeed, our model, derived from a large set of molecules allows to simulate the lipophilicity of structurally diverse molecules. In addition, it compares favorably with this recently published by Schaper and Samitier [26]. Our model has been implemented in AUTOLOGP™ (Version 4.0), a user-friendly software running on IBM™ and compatible PC under Windows™ 3.1 and Windows™ 95 [35].

References

[1] Rekker R.F., Mannhold R., Calculation of Drug Lipophilicity. The Hydrophobic Fragmental Constant Approach, VCH, Weinheim, 1992.

[2] Hansch C., Leo A., Exploring QSAR. Fundamentals and Applications in Chemistry and Biology, ACS Professional Reference Book, American Chemical Society, Washington, DC, 1995.

[3] Hopfinger A.J., Battershell R.D., J. Med. Chem. 19 (1976) 569-573.

[4] Klopman G., Iroff L.D., J. Comput. Chem. 2 (1981) 157-160.

[5] Kasai K., Tomonaga A., Estimation of log *P* value using the physicochemical parameters derived from molecular structure. In: Seydel J.K. (Ed.), QSAR and Strategies in the Design of Bioactive Compounds, VCH, Weinheim, 1985, pp. 277-280.

[6] Reynolds C.H., J. Chem. Inf. Comput. Sci. 35 (1995) 738-742.

[7] Balaban A.T., From Chemical Topology to Three-Dimensional Geometry, Plenum Press, New York, 1997.

[8] Basak S.C., Niemi G.J., Veith G.D., J. Math. Chem. 4 (1990) 185-205.

[9] Niemi G.J., Basak S.C., Veith G.D., Grunwald G., Environ. Toxicol. Chem. 11 (1992) 893-900.

[10] Basak S.C., Gute B.D., Grunwald G.D., J. Chem. Inf. Comput. Sci. 36 (1996) 1054-1060.

[11] Basak S.C., Grunwald G.D., Niemi G.J., Use of graph-theoretic and geometrical molecular descriptors in structure-activity relationships. In: Balaban A.T. (Ed.), From Chemical Topology to Three-Dimensional Geometry, Plenum Press, New York, 1997, pp. 73-116.

[12] Moreau G., Broto P., Nouv. J. Chim. 4 (1980) 359-360.

[13] Moreau G., Broto P., Nouv. J. Chim. 4 (1980) 757-764.

[14] Broto P., Moreau G., Vandycke C., Eur. J. Med. Chem. Chim. Ther. 19 (1984) 66-70.

[15] Chastrette M., Tiyal F., Peyraud J.F., C. R. Acad. Sci. Paris Ser. II 311 (1990) 1057-1060.

[16] Devillers J., Domine D., Chastrette M., A new method of computing the octanol/water partition coefficient. In: Proceedings of QSAR92, July 19-23, 1992, Duluth, MN, USA, p.12.

[17] Devillers J., Domine D., Karcher W., SAR QSAR Environ. Res. 3 (1995) 301-306.

[18] Devillers J., Domine D., Karcher W., Calculating *n*-octanol/water partition coefficients with AUTOLOGP. In: Sanz F., Giraldo J., Manaut F. (Eds.), QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications, J.R. Prous, Barcelona, 1995, pp. 42-44.

[19] Devillers J., Domine D., Karcher W., Polycyclic Arom. Comp. 11 (1996) 211-217.

[20] Domine D., Devillers J., Karcher W., AUTOLOGP versus neural network estimation of *n*-octanol/water partition coefficients. In: Devillers J. (Ed.), Neural Networks in QSAR and Drug Design, Academic Press, London, 1996, pp. 47-63.

[21] Broto P., Devillers J., Autocorrelation of properties distributed on molecular graphs. In: Karcher W., Devillers J. (Eds.), Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology, Kluwer Academic Publishers, Dordrecht, 1990, pp. 105-127.

[22] Hansch C., Leo A., Substituent Constants for Correlation Analysis in Chemistry and Biology, Wiley, New York, 1979.

[23] Eberhart R.C., Dobbins R.W., Neural Network PC Tools: A Practical Guide, Academic Press, San Diego, 1990.

[24] Devillers J., Strengths and weaknesses of the backpropagation neural network in QSAR and QSPR studies. In: Devillers J. (Ed.), Neural Networks in QSAR and Drug Design, Academic Press, London, 1996.

[25] Hansen L.K., Salamon P., IEEE Trans. Pattern Anal. Machine Intell. 12 (1990) 993-1001.

[26] Schaper K.J., Samitier M.L.R., Quant. Struct. Act. Relat. 16 (1997) 224-230.

[27] Hansch C., Leo A., Hoekman D., Exploring QSAR. Hydrophobic, Electronic, and Steric Constants, ACS Professional Reference Book, American Chemical Society, Washington, DC, 1995.

[28] Devillers J., Chambon P., Zakarya D., Chastrette M., Chemosphere 15 (1986) 993-1002.

[29] Devillers J., Chambon P., Zakarya D., Chastrette M., C. R. Acad. Sci. Paris Ser. III 303 (1986) 613-616.

[30] Devillers J., Chambon P., Zakarya D., Chastrette M., Chambon R., Chemosphere 16 (1987) 1149-1163.

[31] Devillers J., Zakarya D., Chambon P., Chastrette M., C. R. Acad. Sci. Paris Ser. III 304 (1987) 195-198.

[32] Devillers J., Zakarya D., Chastrette M., Chemosphere 17 (1988) 1531-1537.

[33] Zakarya D., Belkhadir M., Fkih-Tetouani S., SAR QSAR Environ. Res. 1 (1993) 21-27.

[34] Devillers J., Bintein S., Domine D., Karcher W., SAR QSAR Environ. Res. 4 (1996) 29-38.

[35] AUTOLOGP™ (Version 4.0), CTIS, 3 chemin de la Gravière, 69140 Rillieux La Pape, France.