Quantitative Structure-Property Relationship Analysis for Optical Limiting of Organic Compounds Based on Genetic Algorithm Multivariate Linear Regression

Jie Xu[1,*], Hui Zhang[1], Luoxin Wang[1], Weilin Xu[1], Changhai Yi[1], Hao Liang[2] and Zengchang Li[1]

[1]Key Laboratory of Green Processing & Functional Textiles of New Textile Materials, Ministry of Education, Wuhan Textile University, 430073 Wuhan, Hubei, P.R. China
[2]Department of Chemical Engineering, Huizhou University, 516001 Huizhou, Guangdong, P.R. China

*Corresponding author: Tel:/Fax: +86 27 87426559; E-mail: xujiem@wtu.edu.cn

The quantitative structure-property relationship (QSPR) treatment for optical limiting responses of organic compounds was performed by means of a genetic algorithm based multivariate linear regression (GA-MLR). A four-parameter model ($R^2 = 0.955$ and s = 2.28) was established to predict the output energy (I_{out}). By this correlation, the optical limiting responses could be calculated with satisfactory accuracy. All descriptors involved in the model were derived directly from the molecular structures.

Key Words: Optical limiting, Quantitative structure-property relationship, Genetic algorithm multivariate linear regression.

INTRODUCTION

Optical limiting (OL) phenomenon, a kind of non-linear optical effect, is attracting much attention in the protection of human eyes and optical sensors against high-power pulsed lasers with the development in the laser technology[1]. Several mechanisms could lead to optical limiting behaviour, such as reverse saturable absorption (RSA), two-photon absorption (TPA), non-linear refraction and optically induced scattering[2]. Since the first discovery of optical limiting phenomenon, much work has been done in the exploration of designing and synthesizing materials with excellent optical limiting properties[3-6]. Methods for predicting optical limiting properties of compounds from their molecular structures undoubtedly would be valuable in the search for compounds and materials showing desired optical limiting effects. In the past, many attempts have been made to calculate the second-order hyperpolarizability γ[7-10], but the relation between γ and optical limiting properties of organic materials is not straightforward[11].

The quantitative structure-property relationship (QSPR) approach has become very useful in the prediction of physical and chemical properties of organic compounds. This approach is based on the assumption that the variation of the behaviour of the compounds, as expressed by any measured physico-chemical properties, can be correlated with changes in molecular features of the compounds termed descriptors[12-14]. The advantage of this approach lies in the fact that it requires only the knowledge of the chemical structure and is not dependent on any experimental properties. Once a correlation is established, the structure of any number of compounds with desired properties can be predicted. Thus the QSPR approach has the potential to decrease considerably time and effort to discover new molecules and materials. Although the QSPR has been successfully applied to predict many physicochemical properties[9,10,15-21], there have been relatively few attempts to correlate and predict optical limiting responses of organic compounds. Recently, Lind *et al.*[11] used partial least square (PLS) analysis to develop the first QSPR for prediction of optical limiting response of 23 organic compounds, with 6 descriptors involved. From both a statistical and mechanistic viewpoint, inclusion of variables into a QSPR should meet stringent criteria[22]. The ratio of observations to variables should be as high as possible and at least 5:1[23], while the ratio in this model is less than 5:1 ($23:6 \approx 4:1$). From the statistics, if the number of variables is comparable to the number of training patterns, the parameters of the model may become unstable and unlikely to replicate if the study were to be repeated. In this sense, this model is too complex and will meet these problems.

The main steps involved in QSPR include the data collection, molecular geometry optimization, molecular descriptor generation, descriptor selection, model development and validation[24]. The most commonly used descriptor selection method in QSPR studies is the stepwise regression approach, which can run forward or backward. The stepwise techniques, however, have three main disadvantages. First, each choice

heavily affects the following choices, for example, in forward stepwise version once one of the variables has been selected, all the models that do not contain it will never be taken into account. Second, the final results are expressed by a single combination and then no choice is given to the users[25]. Third, the needed time for solving a problem increases intensely with increasing in the number of molecular descriptors.

The genetic algorithm (GA)[26-28] introduced by Holland[29] has been considered superior to other method of variable selection techniques. It is a search paradigm inspired by natural evolution where the variables are represented as genes on a chromosome (model). It is similar to simplex optimization and evolves from a group of random initial models (population) with fitness scores and searches for chromosomes with better fitness functions (response function scores) through natural selection and the genetic operators, mutation and recombination. The natural selection guarantees the propagation of chromosomes with better fitness in future populations. The genetic algorithm combines genes from two parent chromosomes using the genetic recombination operator to form two new chromosomes (children) that have a high probability of having better fitness than their parents and also explores new response surface (local optima) through mutation. The genetic algorithms offering a combination of hill-climbing ability (natural selection) and a stochastic method (recombination and mutation) are very flexible because they optimize on a representation of variables, not the variables themselves. In addition, the genetic algorithms provide efficient optimization as they use implicit parallelism to process information quickly and require fewer response function evaluations than other automated numerical optimization algorithms.

The aim of the present study is to establish a quantitative structure-property relationship (QSPR) model that could predict optical limiting responses of organic compounds from their molecular structures alone and to discover the main structural features related to the optical limiting responses. The same data set as Lind *et al.*[11] was used and the GA-MLR method was employed to select variables and develop models.

EXPERIMENTAL

The molecular structures of organic compounds (Fig. 1) and the corresponding experimental optical limiting responses (Table-1) were taken from the article by Lind *et al.*[11]. The geometries of the compounds were optimized to ensure that minimum energy conformations were obtained using the Polak-Ribiere algorithm with the semi-empirical AM1 method in Gaussian 03 program package[30]. All optimizations were carried out at a restricted Hartree-Fock level with no configuration interaction. Then totally 721 molecular descriptors for each compound were calculated on the resulting geometry through DRAGON software[31]. These descriptors include Randic molecular profiles, geometrical, RDF (Radial Distribution Function), 3D-MoRSE (3D-Molecule Representation of Structures based on Electron diffraction), WHIM (Weighted Holistic Invariant Molecular) and GETAWAY (GEometry, Topology and Atom-Weights AssemblY) descriptors.
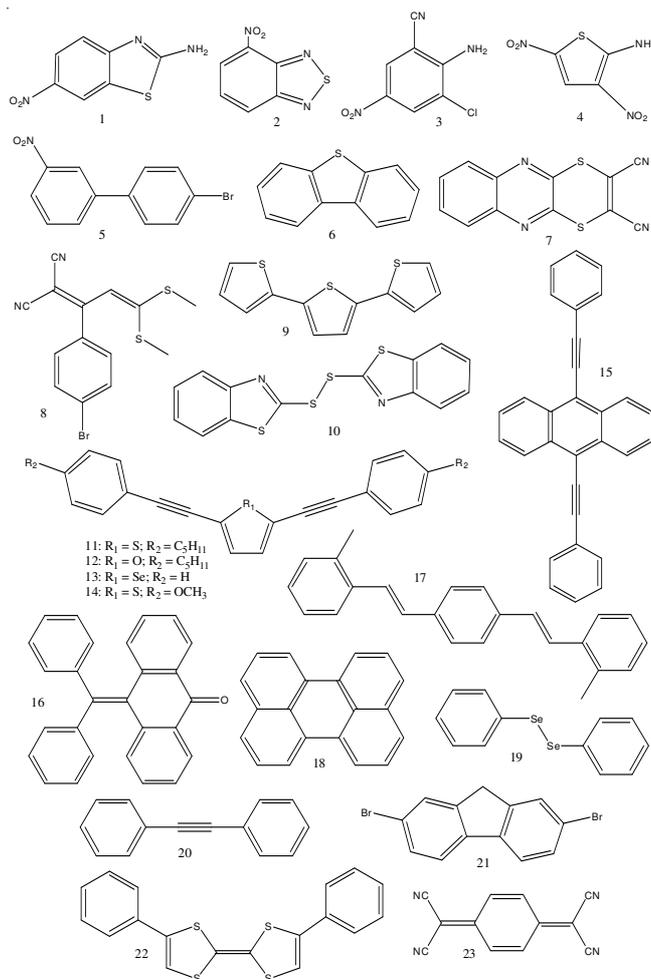

Fig. 1. Compounds' structures included in the data set

TABLE-1 COMPOUNDS USED IN THIS STUDY WITH OBSERVED AND PREDICTED OL RESPONSES				
Obs. T*	Obs. I_{out}**	Pred. I_{out}	Obs. T/I_{out}	Pred. T/I_{out}
99.8	35.7	37.5	0.0280	0.0266
100.0	40.0	37.9	0.0250	0.0264
100.0	33.3	36.4	0.0300	0.0275
86.0	44.0	42.2	0.0195	0.0204
94.3	23.4	20.3	0.0403	0.0465
100.0	36.1	32.5	0.0277	0.0307
96.6	21.9	20.5	0.0441	0.0471
91.9	21.3	23.6	0.0431	0.0390
100.0	23.4	25.0	0.0427	0.0401
99.1	20.5	20.8	0.0483	0.0477
100.0	23.0	24.7	0.0435	0.0405
100.0	20.0	17.4	0.0500	0.0574
100.0	19.8	21.3	0.0505	0.0470
100.0	18.5	21.5	0.0541	0.0466
98.4	7.9	9.0	0.1246	0.1097
97.9	15.8	14.6	0.0620	0.0671
94.1	13.8	13.6	0.0682	0.0689
93.3	11.8	11.7	0.0791	0.0797
93.8	39.9	38.8	0.0235	0.0242
100.0	30.5	33.6	0.0328	0.0297
95.2	20.7	21.7	0.0460	0.0439
82.4	17.7	16.7	0.0466	0.0493
98.0	29.6	27.4	0.0331	0.0357

*Linear transmission (%) at 532 nm, **Output energy (μJ) read at an input energy of 150 μJ, compensated for linear absorption.

In order to reduce redundant and non-useful information, constant or near constant values and descriptors found to be highly correlated pairwise (one of any two descriptors with a correlation greater than 0.99[32]) were excluded in a pre-reduction step. Thus 432 molecular descriptors underwent subsequent descriptor selection.

The GA-MLR was used for variable selection and model development. The size of the population was 50, the probability of crossover 0.5, the probability of mutation 0.1 and the number of evolution generations 200. The quality of the model was scored with the R^2, the adjusted R^2, the leave-one-out (LOO) cross-validated R^2, the standard error of estimation s, the F ratio values and the p values corresponding to 95 % confidence level.

The adjusted R^2 is calculated from

$$R_{adj}^2 = 1 - \left[\left(\frac{N-1}{N-M-1} \right)(1-R^2) \right] \qquad (1)$$

where N = total number of samples in the training set and M = number of descriptors involved in the correlation. The adjusted R^2 is a better measure of the proportion of variance in the data explained by the correlation than R^2 (especially for correlations developed using small datasets) because R^2 is somewhat sensitive to changes in N and M. The adjusted R^2 corrects for the artificiality introduced when M approaches N through the use of a penalty function which scales the result.

A variance inflation factor (VIF) was calculated to test if multicollinearities existed among the descriptors, which is defined as

$$VIF = \frac{1}{1 - R_j^2} \qquad (2)$$

where R_j^2 is the squared correlation coefficient between the jth coefficient regressed against all the other descriptors in the model. Models would not be accepted if they contain descriptors with variance inflation factor above a value of five[33].

RESULTS AND DISCUSSION

Genetic algorithm based multivariate linear regression (GA-MLR) was employed to select variables and develop models, relying on the evolutionary operations of "crossover and mutation" to select optimal combinations of descriptors capable of explaining property variation among the data set. It is clear that univariant correlations between I_{out} and different descriptors have a small value for the correlation coefficient, which indicates that I_{out} is not linearly correlated with any of the molecular descriptors. From a statistical viewpoint the ratio of the number of samples (N) to the number of descriptors (M) should not be too low. Usually, it is recommended that N/M ≥ 5. In the situation of this work, with 23 samples, four descriptors were selected. The final correlation equation is the following:

$$I_{out} = 41.018 - 0.799 \times RDF025v - 11.680 \times Mor27u$$
$$- 12.824 \times H3m + 10.546 \times R5e \qquad (3)$$

N = 23, R = 0.977, R^2 = 0.955, R_{adj}^2 = 0.945, R_{CV}^2 = 0.923, s = 2.28, F = 94.633, p = 0. 000 000

Here, RDF025v is the radial distribution function - 2.5/ weighted by atomic van der Waals volumes[34]; Mor27u is the

3D-MoRSE signal 27/unweighted[35,36]; H3m is the H autocorrelation of lag 3/weighted by atomic masses[37,38] and R5e is the R autocorrelation of lag 5/weighted by atomic Sanderson electronegativities[37,38], respectively (Table-2).

TABLE-2
STATISTICAL CHARACTERISTICS OF DESCRIPTORS IN THE FINAL MODEL

Descriptor	Descriptor type	X	DX	t-Test	p-Value	VIF
Constant		41.018	1.709	23.999	0.000 000	
RDF025v	RDF descriptors	-0.799	0.083	-9.635	0.000 000	1.685
Mor27u	3D-MoRSE descriptors	-11.680	2.130	-5.483	0.000 033	1.709
H3m	GETAWAY descriptors	-12.824	1.381	-9.287	0.000 000	1.084
R5e	GETAWAY descriptors	10.546	1.700	6.202	0.000 007	1.206

In general, the larger the magnitude of the F ratio, the better the model predicts the property values in the training set. The large F ratio of 94.633 indicates that eqn. 3 does an excellent job of predicting the I_{out} values in the training set. Eqn. 3 has an adjusted R^2 value of 0.945, which indicates good agreement between the correlation and the variation in the data. The cross-validated correlation coefficient R_{CV}^2 = 0.923 illustrates the stability of the model by focusing on the sensitivity of the model to the elimination of any single data point. The characteristics and interactions of the four descriptors are given in Tables 2 and 3. The t-values indicate that all the descriptors are highly significant. The variance inflation factor values and the interactions suggest that these descriptors are weakly correlated with each other. Thus, the model can be regarded as an optimal regression equation. The predicted results for the I_{out} values from eqn. 3 are shown in Table-1 and Fig. 2. The calculated optical limiting responses (T/I_{out} values) are also given in Table-1 and Fig. 3. The goodness-of-fit parameters for the T/I_{out} values are R^2 = 0.960 and R_{CV}^2 = 0.931, respectively, which are much better than the corresponding values of the model derived by Lind *et al.*[11] (R^2 = 0.557 and R_{CV}^2 = 0.454), indicating that the data set is described relatively well by these four descriptors and that this model can be expected to be a better predictor of optical limiting responses of organic compounds. Randomization experiments were performed to prove the possible existence of fortuitous correlations. Only low R^2 and R_{CV}^2 values were obtained, indicating that the good results of the original model are not due to a chance correlation or structural dependency of the data set.

TABLE-3
INTERACTIONS BETWEEN DESCRIPTORS AND I_{out}

	RDF025v	Mor27u	H3m	R5e	I_{out}
RDF025v	1				
Mor27u	0.536	1			
H3m	0.022	-0.222	1		
R5e	0.234	-0.162	0.119	1	
I_{out}	-0.752	-0.644	-0.379	0.196	1

By interpreting the descriptors involved in eqn. 3, it is possible to gain some insights into the factors that may affect
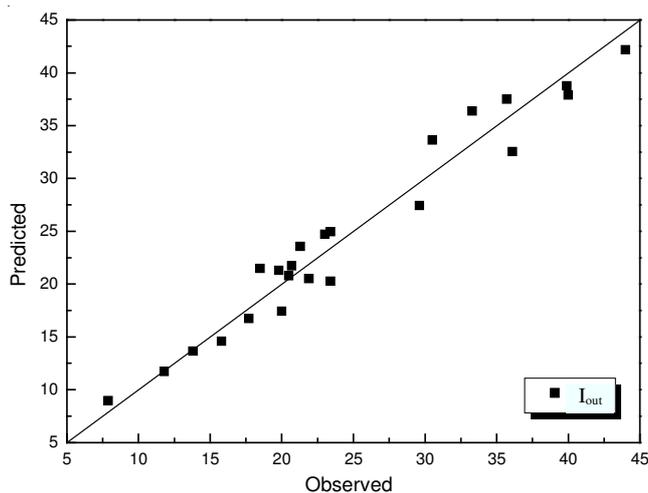
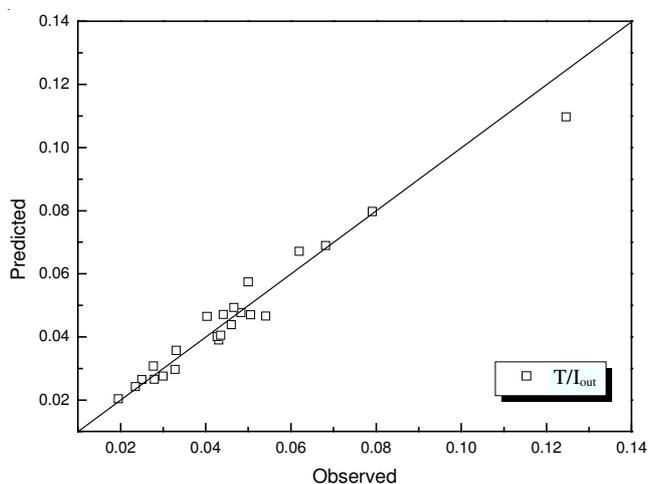Fig. 2.   Predicted *versus* observed values of I_{out} with the final model ($R^2 = 0.955$)

Fig. 3.   Predicted T/I_{out} values (calculated from predicted I_{out} values) *versus* observed values of T/I_{out} ($R^2 = 0.960$)

the I_{out} values. According to the t-test (Table-2), the most significant descriptor appearing in the QSPR model is the descriptor RDF025v, which decreases the I_{out} values. RDF025v is one of the RDF descriptors which have recently been proposed based on a radial distribution function. The RDF descriptors can be interpreted as the probability distribution of finding an atom in a spherical volume of radius r. The general form of the radial distribution function is represented by:

$$RDFrw = f \cdot \sum_{i=1}^{nAT-1} \sum_{j=i+1}^{nAT} w_i \cdot w_j \cdot e^{-\beta(r-r_{ij})^2} \qquad (4)$$

where f is a scaling factor (assumed equal to one in the calculations), w_i and w_j are characteristic properties of the atoms i and j (including unweighted, masses, Van Der Waals volumes, Sanderson electronegativities and polarizabilities), r_{ij} = interatomic distance and nAT = number of atoms in the molecule[34]. RDFrw is generally calculated at a number of discrete points with defined intervals. Besides information about interatomic distances in the entire molecule, RDF25v provides further information about atomic van der Waals volumes. The negative sign of RDF25v indicates that organic compounds with greater volumes would have smaller I_{out} values

and thus higher optical limiting responses. The contributions of this descriptor in optical limiting responses are in agreement with the contributions that one would expect for the conjugation length of molecules as larger organic compounds with optical limiting properties usually exhibit longer conjugation.

The second significant descriptor in the QSPR model is Mor27u, which receives a negative regression coefficient. Mor27u is a 3D representation of molecular structures based on electron diffraction descriptor (3D-MoRSE descriptor)[35,36] which is calculated by summing atomic weights viewed by a different angular scattering function. The values of these descriptor functions are calculated at 32 evenly distributed values of scattering angle(s) in the range of 0-31 Å^{-1} from the three dimensional atomic coordinates of a molecule. The 3D-MoRSE descriptor is calculated using following expression

$$Morsw = \sum_{i=1}^{nAT-1} \sum_{j=i+1}^{nAT} w_i w_j \frac{\sin(s-r_{ij})}{s \cdot r_{ij}} \qquad (5)$$

where nAT is the number of atoms, s = scattering angle, r_{ij} is the interatomic distance, w_i and w_j are atomic properties of ith and jth atom, respectively. The value and sign of the 3D-MoRSE descriptor depend, to a large extent, on the values of s and r_{ij}. Thus, it could not be concluded that certain descriptor has a specific effect on I_{out} values, either negative or positive, only taking into account the coefficient sign of the descriptor in the present model. When the coefficient and the descriptor have the same sign, the contribution of the descriptor is positive, else, negative.

There are two GETAWAY descriptors (H3m and R5e) appearing in the QSPR model. The GETAWAY descriptors[37,38] have recently been proposed as chemical structure descriptors derived from a new representation of molecular structure, the molecular influence matrix. These descriptors, as based on spatial autocorrelation, encode information on the effective position of substituents and fragments in the molecular space. Moreover, they are independent of molecule alignment and, to some extent, account also for information on molecular size and shape as well as for specific atomic properties. H3m and R5e are calculated by eqns. 6 and 7, respectively, where nAT = number of atoms, h_{ii} and h_{jj} are the leverages of ith and jth atom, m_i and m_j are their atomic masses, e_i and e_j are their atomic Sanderson electronegativities, r_{ij} = interatomic distance, d_{ij} = topological distance, $\delta(3; d_{ij}; h_{ij})$ is a Dirac-δ function (δ = 1 if d_{ij} = 3 and h_{ij} > 0, zero otherwise) and $\delta(5, d_{ij})$ is a Dirac-δ function (δ = 1 if d_{ij} = 5, zero otherwise).

$$H3m = \sum_{i=1}^{nAT-1} \sum_{j>1} h_{ij} \cdot m_i \cdot m_j \cdot \delta(3; d_{ij}; h_{ij}) \qquad (6)$$

$$R5e = \sum_{i=1}^{nAT-1} \sum_{j>1} \frac{\sqrt{h_{ii} \cdot h_{jj}}}{r_{ij}} \cdot e_i \cdot e_j \cdot \delta(5, d_{ij}) \qquad (7)$$

The negative sign of H3m in the QSPR model indicates that compounds with larger masses would have smaller I_{out} values and thus higher optical limiting responses. The coefficient of R5e in eqn.3 is positive, meaning that compounds containing atoms with greater electronegativities would possess larger I_{out} and thus lower optical limiting responses.

Conclusion

In this paper, a successful quantitative structure-property relationship model was reported for the prediction of optical limiting responses of organic compounds. The R^2 of the correlation for T/I_{out} values was 0.960, the R_{CV}^2 was 0.931 and the standard error was 0.00462. The model presented here relies solely on descriptors derived from the molecular structures and thus it is applicable to regular organic compounds. Therefore, this quantitative structure-property relationship model should be useful in the development of new optical limiting compounds.

REFERENCES

1. M. Hanack, D. Dini, M. Barthel and S. Vagin, *Chem. Record.*, **2**, 129 (2002).
2. G.S. He, G.C. Xu, P.N. Prasad, B.A. Reinhardt, J.C. Bhatt and A.G. Dillard, *Opt. Lett.*, **20**, 435 (1995).
3. F. Guo, B. Guo, Z. Hu, B. Chen and Q. Zhang, *J. Optoelectron. Adv. Mater.*, **9**, 2152 (2007).
4. F. Guo, X. Sun, H. Liang, Z. Hu, Q. Zhang and H. Ming, *Chin. J. Chem. Phys.*, **21**, 227 (2008).
5. X. Su, H. Xu, Y. Deng, J. Li, W. Zhang and P. Wang, *Mater. Lett.*, **62**, 3818 (2008)
6. C. Girardot, B. Cao, J.-C. Mulatier, P.L. Baldeck, J. Chauvin, D. Riehl, J.A. Delaire, C. Andraud and G. Lemercier, *Chem. Phys. Chem.*, **9**, 1531 (2008).
7. S. Yamada, M. Nakano and K. Yamaguchi, *Chem. Phys. Lett.*, **276**, 375 (1997).
8. M. Yang and Y. Jiang, *Chem. Phys.*, **274**, 121 (2001).
9. K. Oberg, A. Berglund, U. Edlund and B. Eliasson, *J. Chem. Inf. Comp. Sci.*, **41**, 811 (2001).
10. J. Xu, B. Chen, W. Xu, S. Zhao, C. Yi and W. Cui, *Chemometr. Intell. Lab. Syst.*, **87**, 275 (2007).
11. P. Lind, C. Lopes, K. Öberg and B. Eliasson, *Chem. Phys. Lett.*, **387**, 238 (2004).
12. X.J. Yao, Y.W. Wang, X.Y. Zhang, R.S. Zhang, M.C. Liu, Z.D. Hu and B.T. Fan, *Chemometr. Intell. Lab. Syst.*, **62**, 217 (2002).
13. J. Xu, B. Guo, B. Chen and Q. Zhang, *J. Mol. Model.*, **12**, 65 (2005).
14. J. Xu, Z. Zheng, B. Chen and Q. Zhang, *QSAR Comb. Sci.*, **25**, 372 (2006).
15. J. Xu, B. Chen, Q. Zhang and B. Guo, *Polymer*, **45**, 8651 (2004).
16. J. Xu, B. Chen and H. Liang, *Macromol. Theory Simul.*, **17**, 109 (2008).
17. J. Xu, H. Liang, B. Chen, W. Xu, X. Shen and H. Liu, *Chemometr. Intell. Lab. Syst.*, **92**, 152 (2008).
18. X. Yu, X. Wang, H. Wang, A. Liu and C. Zhang, *J. Mol. Struct. (Theochem.)*, **766**, 113 (2006).
19. L.J.D.O. Figueiredo and F.M.S. Garridob, *J. Mol. Struct. (Theochem.)*, **539**, 75 (2001).
20. X.D. Zeng, X. Xu, B.F. Wang and B.C. Wang, *Chin. Chem. Lett.*, **15**, 753 (2004).
21. J. Xu and B. Chen, *J. Mol. Model*, **12**, 24 (2005).
22. M.T.D. Cronin and T.W. Schultz, *J. Mol. Struct. (Theochem.)*, **622**, 39 (2003).
23. J.G. Topliss and R.J. Costello, *J. Med. Chem.*, **15**, 1066 (1972).
24. A. Yasri and D. Hartsough, *J. Chem. Inf. Comp. Sci.*, **41**, 1218 (2001).
25. J. Devillers, Genetic Algorithms in Molecular Modeling, Academic Press, London (1996).
26. H. Kubinyi, *Quant. Struct.-Act. Rel. Pharmacol. Chem. Biol.*, **13**, 285 (1994).
27. H. Kubinyi, *Quant. Struct.-Act. Rel. Pharmacol. Chem. Biol.*, **13**, 393 (1994).
28. H. Kubinyi, *J. Chemometrics*, **10**, 119 (1996).
29. J.H. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, Chicago (1975).
30. M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, J.A. Montgomery Jr., T. Vreven, K.N. Kudin, J.C. Burant, J.M. Millam, S.S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G.A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, Y. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, I.E. Knox, H.P. Hratchian, J.B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, P.Y. Ayala, K. Morokuma, G.A. Voth, P. Salvador, J.J. Dannenberg, V.G. Zakrzewski, S. Dapprich, A.D. Daniels, M.C. Strain, O. Farkas, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J.V. Ortiz, Q. Cui, A.G. Baboul, S. Clifford, J. Cioslowski, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, M. Challacombe, P.M.W. Gill, B. Johnson, W. Chen, M.W. Wong, C. Gonzalez and J.A. Pople, Gaussian 03, Revision E.01; Gaussian Inc.: Wallingford CT (2004).
31. R. Todeschini, V. Consonni, A. Mauri and M. Pavan, DRAGON for Windows (Software for Molecular Descriptor Calculations), Version 5.4; TALETE srl: Milan (2006).
32. H. Liu and P. Gramatica, *Bioorg. Med. Chem.*, **15**, 5251 (2007).
33. A.J. Holder, D.M. Yourtee, D.A. White, A.G. Glaros and R. Smith, *J. Comp. Aid. Mol. Des.*, **17**, 223 (2003).
34. M.C. Hemmer, V. Steinhauer and J. Gasteiger, *Vibrat. Spect.*, **19**, 151 (1999).
35. J. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, L. Steinhauer and V. Steinhauer, *J. Chem. Inf. Comp. Sci.*, **36**, 1030 (1996).
36. J. Schuur, P. Selzer and J. Gasteiger, *J. Chem. Inf. Comp. Sci.*, **36**, 334 (1996).
37. V. Consonni, R. Todeschini and M. Pavan, *J. Chem. Inf. Comp. Sci.*, **42**, 682 (2002).
38. V. Consonni, R. Todeschini, M. Pavan and P. Gramatica, *J. Chem. Inf. Comp. Sci.*, **42**, 693 (2002).