Application of quantum topological molecular similarity descriptors in QSPR study of the O-methylation of substituted phenols

ARTICLE *in* JOURNAL OF COMPUTATIONAL CHEMISTRY · JANUARY 2008

Impact Factor: 3.59 · DOI: 10.1002/jcc.20787 · Source: PubMed

2 AUTHORS:

Bahram Hemmateenejad
Shiraz University
183 PUBLICATIONS **2,794** CITATIONS

SEE PROFILE

Afshan Mohajeri
Shiraz University
68 PUBLICATIONS **556** CITATIONS

SEE PROFILE

Application of Quantum Topological Molecular Similarity Descriptors in QSPR Study of the O-Methylation of Substituted Phenols

BAHRAM HEMMATEENEJAD,[1,2] AFSHAN MOHAJERI[1]
[1]*Chemistry Department, Shiraz University, Shiraz, Iran*
[2]*Medicinal and Natural Products Chemistry Research Center, Shiraz University of Medical Sciences, Shiraz, Iran*

Received 5 December 2006; Revised 7 April 2007; Accepted 8 May 2007
DOI 10.1002/jcc.20787
Published online 15 June 2007 in Wiley InterScience (www.interscience.wiley.com).

Abstract: The usefulness of a novel type of electronic descriptors called quantum topological molecular similarity (QTMS) indices for describing the quantitative effects of molecular electronic environments on the O-methylation kinetic of substituted phenols has been investigated. QTMS theory produces for each molecule a matrix of descriptors, containing bond (or structure) information in one dimension and electronic effects in another dimension, instead of other methods producing a vector of descriptors for each molecule. A collection of chemometrics tools including principal component analysis (PCA), partial least squares (PLS), and genetic algorithms (GA) were used to model the structure-kinetic data. PCA separated the bond and descriptor effects, and PLS modeled the effects of these parameters on the rate constant data, and GA selected the most relevant subset of variables. The model performances were validated by both cross-validation and external validation. The results indicated that the proposed models could explain about 95% of variances in the rate constant data. The significant effects of variables on the reaction kinetic were identified by calculating variable important in projection (VIP). It was found that the rate constant of esterification of phenols is highly influenced by the electronic properties of the C2—C1—O—H fragment of the parent molecule. Indeed, the C2—X and C4—X bonds (corresponding to ortho and para substituents) were found as highly influential parameters. All of the eight calculated QTMS indices were found significant however, λ_1, λ_2, λ_3, ε, and $K(r)$ were detected as highly influential parameters.

 J Comput Chem 29: 266–274, 2008

Key words: quantum topological molecular similarity; rate constant; phenol; QSPR; O-methylation

Introduction

Some of fundamental thermodynamic, physical, and chemical properties of many compounds are not available in literature and their measurement can be costly and time-consuming. The availability of biological data is even less. Nevertheless, values for such properties are required for the design, control and understanding of chemical processes, the study of environmental behavior, and for the drug design. Thus, reliable and accessible estimation methods are needed.[1]

Quantitative structure property/activity relationships (QSPR/QSAR) studies are connections between the molecular structure of organic compound and their chemical or biological properties.[2–7] The properties have been determined by a regression analysis of that property to aspects of some molecular descriptors. Once a correlation is established, the structure of any number of compounds with desired properties can be predicted and/or a more deep knowledge about the mechanism of the action can be obtained. Molecular descriptors play fundamental roles in the QSAR/QSPR studies and finding new molecular descriptors with higher correlation toward activity/property is in the frontier of QSAR/QSPR researches.[8] Electronic descriptors obtained from quantum chemical calculations have been found major popularity and there is a challenge between calculation complexity and accuracy to select the quantum chemical calculation methods (i.e., semiempirical and *ab initio*).[9] Recent progress in computational hardware and the development of efficient algorithms have assisted the routine development of molecular quantum chemical calculations. Quantum chemical calculations are thus an attractive source of new molecular descriptors, which can, in principle, express all of the electronic and geometric properties of molecules and their interactions.[10–16]

This article contains supplementary material available via the Internet at http://www.interscience.wiley.com/jpages/0192-8651/suppmat

Correspondence to: B. Hemmateenejad; e-mail: hemmatb@sums.ac.ir

Contract/grant sponsor: Shiraz University and Shiraz University of Medical Sciences

Recently, a new class of electronic descriptors called quantum topology molecular similarity (QTMS) indices, introduced by O'Brein and Popelier,[17,18] has been shown to be successful in a verity of QSAR and QSPR calculations. The theory behind the calculation of the QTMS indices uses the idea of theory of "atoms in molecules" (AIM) pioneered by Bader[19] to specify the electronic information in a molecular system. This theory is deeply rooted in quantum mechanics and can be used to enhance chemical insight through *ab initio* wavefunctions. It has been demonstrated that QTMS offers a reliable alternative to electronic parameters and delivers an excellent QSARs of environmental, biological, and industrial interest.[20–26]

In the recent years, phenolic hydroxyl group has received major attention because of its wide range of activity.[27–29] On one hand, it represents antioxidant activity whilst on the other hand, it exhibits significant toxicity. Indeed, predicting the rate constants of phenolic compounds by computational methods are important not only in chemical aspects but also for environmental, biological, and industrial applications. QTMS indices have been previously used by Popelier and coworkers to assess the toxicity of chlorophenols[26] as well as the cytotoxicity of ortho alkyl-substituted phenols.[25] In addition, Chaudry and Popelier developed predictive QSPR models for hydrolysis rate constant of esters using QTMS indices.[22] The rate constant of the O-methylation (or esterification) reaction of phenolic compounds, the subject of this article, were studied by Cork and Hayashi.[30] They found a linear correlation model between the rate constant and the semiempirical calculated charge density on the phenolic oxygen. In this article, we will report the results of our investigation using QTMS indices to account the quantitative effects of structural variations on the kinetic of O-methylation reaction of substituted phenol via QSPR analyses. The separate effects of QTMS descriptors and each chemical bond of the phenol derivatives will be described.

Methods

Computational Details

The theoretical bases of QTMS indices have been extensively described by Popelier and coworkers,[20–26] and herein we have described them briefly. The theory of AIM takes advantage of the electron density as an information source from which to formulate chemical concepts. The properties of a molecular charge distribution are summarized in terms of its critical points. This corresponds to the points where the gradient of electron density vanishes ($\bigtriangledown\rho = 0$). A critical point is characterized by the sign of its three principal curvatures of ρ. A bond critical point (BCP), with one positive and two negative curvatures, is located between two atomic centers and denotes the presence of a bond. Another useful quantity to characterize a bond is the Hessian of the charge density. The Hessian is a matrix describing all possible second derivatives of ρ with respect to coordinates. Diagonalization of the Hessian yields three eigenvalues, λ_1, λ_2, and λ_3. The Laplacian of density, $\bigtriangledown^2\rho$, at the BCP is the sum of its three principal curvatures at each point in space:

$$\nabla^2\rho = \lambda_1 + \lambda_2 + \lambda_3 \tag{1}$$

This provides a measure of the extent to which the charge density is locally compressed or expanded in a bond.

Another quantity derived from the Hessian eigenvalues is the ellipticity of a bond (ε) at the BCP and is defined as:

$$\varepsilon = \frac{\lambda_1}{\lambda_2} - 1 \tag{2}$$

Ellipticity provides the measure of the extent to which charge is accumulated in a given plane and can be used as a quantitative index of the π-character of a bond.[20] Bonds can further be characterized by evaluating the two types of kinetic energy densities denoted by Lagrangian kinetic energy, $G(r)$,

$$G(r) = 1/2N \int d\tau' \nabla\psi^* \nabla\psi \tag{3}$$

and Hamiltonian kinetic energy, $K(r)$,

$$K(r) = -1/4N \int d\tau' [\psi^* \nabla^2 \psi + \psi \nabla^2 \psi^*] \tag{4}$$

where N is the number of electrons and $N\int d\tau'$ summarizes the one-electron integration mode. The quantities introduced above can be used together as chemical descriptors for a bond.

Calculation of QTMS indices is preceded in two steps. In the first step, a geometry optimization is performed for each molecule to obtain structural parameters and wave functions at a level of theory. Thus, the molecules are optimized at the semiempirical PM3 method, which are followed by a density functional B3LYP/6-311++G** calculation. All calculations were carried out using GAUSSIAN 98 program.[31] In the second step, wave functions calculated at the B3LYP/6-311++G** were used for the topological analyses of the electron densities. The AIM2000 program[32] was employed for calculating the BCPs and visualizing the bond paths. In this step, the BCPs are located for each individual bond in the molecule. Analysis of the electron densities produce eight properties (ρ, $\bigtriangledown^2\rho$, λ_1, λ_2, λ_3, ε, $G(r)$, and $K(r)$) for each BCP. Each property can be used as a descriptor variable. Thus, for each molecule, a matrix of QTMS descriptor with dimension of ($8 \times n$) is obtained, where n is the number of chemical bonds in the basic molecular skeleton.

Model Development

The rate constant data of the esterification of the phenol derivatives were taken from the work of Cork and Hayashi[30] (Table 1). The data were transferred to the logarithmic scale as $\log(k_X/k_H)$, where k_X is the first-order rate constant of the substituted phenol and k_H is the corresponding value for the parent molecule (i.e., phenol). A total of 36 phenol derivatives used in this study (Table 1), among which 10 molecules were randomly selected as prediction set and the rest was used in model development step. As it was noted previously, QTMS theory produces a matrix of descriptors for each molecule, whose number of rows and columns equal to the number of bond for the basic

Table 1. Chemical Structure of the Phenol Derivatives Used in This Study and Their Corresponding Experimental and Cross-Validation Predicted Rate Constant Data.

						$\log(k_X/k_H)$	
No.	X2	X3	X4	X5	X6	Experimental	Predicted by GA-PLS
1	H	H	H	H	H	0.000	0.021
2	H	H	Me	H	H	0.216	0.223
3	H	H	OMe	H	H	0.226	0.142
4	H	H	F	H	H	−0.001	−0.263
5	H	H	Cl	H	H	−0.111	−0.473
6	H	H	COMe	H	H	−1.246	−1.453
7	H	H	$CONH_2$	H	H	−1.629	−1.321
8	H	H	CN	H	H	−1.578	−1.335
9	H	H	NO_2	H	H	−2.687	−2.519
10	H	Me	H	H	H	0.412	0.012
11	H	OMe	H	H	H	0.160	0.055
12	H	F	H	H	H	−0.663	−0.968
13	H	Cl	H	H	H	−0.518	−0.346
14	H	COMe	H	H	H	−0.426	−0.732
15	H	CN	H	H	H	−0.956	−0.477
16	H	NO_2	H	H	H	−0.996	−1.233
17	Me	H	H	H	H	0.164	−0.083
18	OMe	H	H	H	H	−0.490	−0.515
19	F	H	H	H	H	−0.396	0.281
20	Cl	H	H	H	H	−0.611	−0.922
21	COMe	H	H	H	H	−1.863	−1.305
22	$CONH_2$	H	H	H	H	−1.629	−0.680
23	CN	H	H	H	H	−1.687	−1.618
24	NO_2	H	H	H	H	−2.930	−3.194
25	H	Me	H	Me	H	0.200	−0.017
26	H	OMe	H	OMe	H	0.267	0.359
27	H	OMe	H	Cl	H	−0.420	−0.369
28	H	Cl	H	Cl	H	−1.164	−1.007
29	H	Me	NO_2	H	H	−2.518	−2.265
30	OMe	H	Me	H	H	−0.175	−0.259
31	Cl	H	Cl	H	H	−1.196	−0.685
32	OMe	H	COMe	H	H	−1.554	−1.778
33	OMe	H	NO_2	H	H	−3.356	−3.297
34	NO_2	H	NO_2	H	H	−5.189	−5.396
35	$CONH_2$	H	Cl	H	H	−2.231	−2.425
36	$CONH_2$	OH	Cl	H	Cl	−3.146	−3.317

molecular structure and number of QTMS parameters, respectively. In this work, the structural backbone of the phenolic compounds (Fig. 1) contains 13 bonds (i.e., 6 C—C bonds, 5 C—H, or C—X bonds, one O—H bond, and one C—O bond). Indeed, eight QTMS parameters were calculated for each bond. Therefore, each QTMS descriptor data matrix has a dimension of (13 × 8) and the total number of calculated descriptors for each molecule is equal to 104. Thus, the descriptor data matrix of entire set of molecules has a dimension of (36 × 104).

Modeling of $\log(k_X/k_H)$ as a function of QTMS indices was performed by partial least square regression.[33] The input (or predictor variables) of the PLS model was the original QTMS indices unfolded to a row vector for each molecule or was the first principal component of the QTMS descriptor data matrix of each molecule. In the latter case, the matrix of QTMS indices of an individual molecule was subjected to principal component analysis (PCA)[33] and the first principal component (or eigenvector) of this matrix was considered as the input variables of that molecule for PLS analysis. In each case, the PLS regression was achieved by the entire set of input variables as well as by the subset of variables selected by genetic algorithm (GA).[34] The model refinement procedure was the use of leave-one-out cross-validation (LOO-CV) method to select the optimum number of PLS latent variables.[35]

The GA used here was the same as we used previously.[34,35] The GA used a binary bit string representation as the coding technique for a given problem; the presence or absence of a descriptor in a chromosome was coded by 1 or 0. A string was composed of several genes that represented a specific characteristic that should be studied. In the present case, a string was composed of 76 genes, representing the presence or absence of a descriptor (by 1 or 0, respectively). By encoding various variables with bit strings, called chromosomes, the initial population was created randomly. The population size was varied between 50 and 250 for different GA runs. Besides, the number of genes with the values of 1 was kept relatively low to have a small subset of descriptors, i.e., the probability of generating 0 for a gene was set greater (at least 65%). Using the initial population, PLS regression models (using different subset of descriptors) were built and their predictivity was evaluated by cross-validation. The chromosomes with the least numbers of selected descriptors and the highest fitness were marked as informative chromosomes. These chromosomes were kept for natural selecting and crossovering steps to survive in the next generation preferentially. For a typical run, the evolution of the generation was terminated when 90% of the generations reached to the same fitness. GA-PLS was run many times each with different GA parameters such as mutation and crossovering rates and population size.

The subroutines for doing PCA and PLS were written in MATLAB (Mathwork, version 7). For variable selection by GA, the PLS toolbox developed by eigenvector company (Eigenvector Research) was employed. A Pentium IV personal computer with windows XP operating system was used throughout.

Results and Discussion

The rate constant data, shown in Table 1, reveal a dramatic change in the O-methylation reaction rate of phenol derivatives

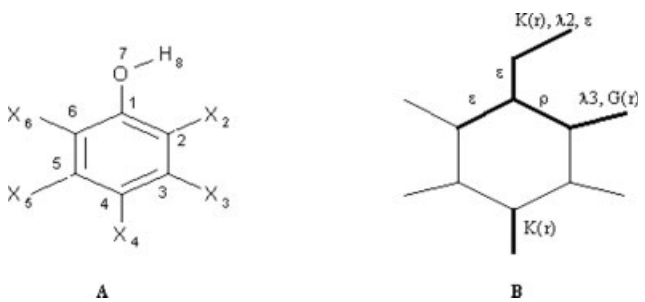

Figure 1. A: structure and numbering of the basic skeleton of the phenol derivatives. B: The chemical bonds of phenols that are detected as highly influential on the rate constant by GA-PLS.

Table 2. List of Collinear Descriptors.

No.	Collinear descriptors
1	$K(r)_{C1-C2}$, $\rho_{\mathbf{C1-C2}}$
2	$\nabla^2\rho_{C1-C2}$, $\nabla^2\rho_{C2-C3}$, $\nabla^2\rho_{C4-C5}$, $\nabla^2\rho_{C6-C1}$, $\nabla^2\rho_{O7-H8}$
3	$G(r)_{C6-X6}$, $K(r)_{C2-C3}$
4	$K(r)_{C3-C4}$, ρ_{C3-C4}
5	$K(r)_{C4-C5}$, ρ_{C4-C5}
6	$\lambda_{1,C5-C6}$, ρ_{C5-C6}, $K(r)_{C5-C6}$
7	$K(r)_{C6-C1}$, ρ_{C6-C1}
8	ρ_{C1-O7}, $\lambda_{1,C1-O7}$, $\lambda_{2,C1-O7}$, $\lambda_{3,C1-O7}$, $K(r)_{C1-O7}$, $G(r)_{C1-O7}$
9	$\lambda_{2,C2-X2}$, $\lambda_{1,C2-X2}$
10	$\lambda_{2,C3-X3}$, $\lambda_{1,C3-X3}$
11	$K(r)_{C3-X3}$, $\lambda_{3,C3-X3}$
12	ρ_{C4-X4}, $\lambda_{1,C4-X4}$, $\lambda_{2,C4-X4}$, $\lambda_{1,C4-X4}$
13	$G(r)_{C4-X4}$, ε_{C4-X4}
14	$\lambda_{2,C5-X12}$, $\lambda_{1,C5-X12}$
15	$K(r)_{C5-X4}$, $\lambda_{3,C5-X4}$
16	ρ_{C6-X6}, $\lambda_{1,C6-X6}$, $\lambda_{2,C6-X6}$, $\lambda_{3,C6-X6}$, $K(r)_{C6-X6}$

by changing substituent from electron-donating to electron-withdrawing groups. Whilst the 3-methyl phenol has the highest rate constant (i.e., 26.4), the lowest rate constant has been reported for 2,4-dintro phenol derivative (i.e., 6.6×10^{-5}). This five-order change in the rate constant suggests that the electronic features of the phenol derivatives play an important role in the mechanism of the O-methylation reaction of phenols. The significant effect of the substituent on the charge density of phenolic oxygen and consequently on the rate constant of O-methylation has been previously reported by Cork and Hayashi.[30] To obtain a more deep insight to the reaction mechanism of O-methylation of substituted phenols and to more understand the effect of electronic features on the rate constant, we used quantum topological molecular similarity (QTMS) descriptors for modeling the rate constant of the O-methylation of phenol derivatives. The rate constant data were converted to the logarithmic scale in relative to the rate constant of phenol as $\log(k_X/k_H)$. Modeling of the $\log(k_X/k_H)$ as a function of QTMS descriptors is proceeded by multiple linear regression (MLR) or partial least squares (PLS) regression.

Collinearity between the variables can introduce instability into QSAR models based on MLR analysis. In addition, if a high correlation exists between the descriptors, it is impossible to delimitate the effects of each descriptor in MLR analysis.[36] Therefore, MLR models should be carried out only using uncorrelated descriptors. Meanwhile, this analysis reduces the dimension of the data set before applying PLS analysis. To do so, correlation between each one of descriptors and with rate constant data was calculated. The detected groups of collinear descriptors are listed in Table 2. The full correlation matrix can be found in the supplementary materials. As it is obvious from Table 2, 16 groups of descriptors have been detected as collinear, and in some groups more than two collinear descriptors are observed. For each subset of collinear descriptors, one of them, which had the highest correlation with rate constant (those bolded in Table 2), was retained and the rest were omitted. By this manner, 28 descriptors were discarded and 76 descriptors were remained for the future analyses.

Both the methods used in this study (i.e., MLR and PLS) need a one-dimensional array of descriptors and therefore it is necessary to convert the two-dimensional array (or matrix) of descriptors to one-dimensional array (or vector). To do so, some different approaches were employed.

At the first, the data were analyzed to find which chemical bonds of the phenol molecular skeleton (Fig. 1) have significant impact on the rate constant of O-methylation. To do so, separate QSPR models were developed for each bond employing QTMS descriptors corresponding to that bond. It should be noted that the collinear descriptors indicated in Table 2 were not considered in the development of the MLR-based QSPR models. As it is observed from Table 2, for some bonds more than two QTMS descriptors have been detected as collinear variables. These bonds are C5—C6, C1—O7, C3—X3, C4—X4, C5—X5, and C6—X6. Interestingly, none of the QTMS indices of O7—H bond were collinear, which means that the QTMS indices of this bond are varied independently when substitution pattern on the phenolic molecular structure are changed. The chemical bonds for which significant QSPR equations are obtained can be considered as molecular fragments whose electronic properties have significant effects on the rate constant. In the case of each bond, the most convenient QSPR equation was obtained by multiple linear regression analysis (MLR) employing stepwise selection of variables. The coefficient of multiple determinations (R^2) was used to measure the goodness of fit for each model. In addition, the resulted models were validated for generalization and predictivity by LOO-CV utilizing the square of correlation coefficient for cross-validation (R^2_{CV}). In addition, the prediction ability of the models was also confirmed by random splitting of the data into calibration (26 molecules) and prediction (10 molecules) sets. The rate constant of the prediction set molecules was calculated based on the model coefficients calculated from the calibration data, and then the correlation coefficient between the predicted and experimental rate constants (R^2_P) were calculated. The resulted QSPR equations obtained for each bond are sum-

Table 3. QSPR Equations Obtained From QTMS Descriptors of Different Bonds.

No.	Bond	Selected QTMS indices	R^2	R^2_{CV}	R^2_P
1	C1—C2	ρ	0.393	0.318	0.464
2	C2—C3	NA	–	–	–
3	C3—C4	$\nabla^2\rho$	0.117	0.067	0.128
4	C4—C5	λ_2	0.156	0.060	0.238
5	C5—C6	λ_1	0.649	0.565	0.715
6	C6—C1	$G(r)$	0.209	0.107	0.185
7	C1—O7	λ_3	0.833	0.802	0.825
8	O7—H8	λ_2, ε	0.906	0.867	0.888
9	C2—X2	ε, $K(r)$	0.548	0.447	0.605
10	C3—X3	NA	–	–	–
11	C4—X4	ε	0.415	0.328	0.589
12	C5—X5	NA	–	–	–
13	C6—X6	NA	–	–	–

NA, not applicable.

Table 4. QSPR Equations Obtained for QTSM From Different Bonds.

No.	QTMS index	Selected bonds	R^2	R^2_{CV}	R^2_P
1	ρ	C3—C4, C5—C6, C6—C1, C1—O7	0.893	0.823	0.860
2	$\bigtriangledown^2\rho$	O7—H8, C6—X6	0.432	0.253	0.271
3	λ_1	C1—C2, C1—O7, O7—H8	0.835	0.714	0.756
4	λ_2	C5—C6, C1—O7, O7—H8, C4—X	0.884	0.778	0.885
5	λ_3	C1—O7, O7—H8	0.866	0.785	0.841
6	ε	C5—C6, O7—H8, C6—X	0.902	0.881	0.909
7	$K(r)$	C1—O7, O7—H8, C4—X	0.906	0.866	0.920
8	$G(r)$	C3—C4, C5—C6, C6—C1, C1—O7	0.902	0.853	0.894

marized in Table 3. As is observed, the QSPR models obtained for C1—O7 and O7—H8 bonds have the highest statistical quality, which confirms the high impact of these bonds in the esterification rate constant of phenol derivatives. This result is not surprising since C1—O7 and O7—H8 bonds are directly involved in the O-methylation reaction. The resulted QSPR equations for C1—C2, C5—C6, C2—X, and C4—X bonds have moderate statistical quality. For bonds which are far from the reaction center, including C3—C4, C4—C5, C6—C1, C2—C3, C3—X, C5—X, and C6—X, bonds no significant correlation equation or equations with very low statistical quality were obtained.

In another trial, we emphasized on the QTMS indices and attempted to find QTMS descriptors played the most significant role in O-methylation kinetic of phenols. In the case of each QTMS index, 13 calculated values belonging to 13 chemical bonds are available. According to the data reported in Table 2, the colinearity between the QTMS of different bond for an individual QTMS descriptor was investigated first. It is obvious that except for $\bigtriangledown^2\rho$ no significant colinearity has been observed for QTMS indices. For $\bigtriangledown^2\rho$ index, its values on five bonds, including C1—C2, C2—C3, C4—C5, C6—C1, and O7—H, represented significant colinearity, where among them O7—H was retained and the rest were not considered in the MLR modeling. Hence, MLR analysis for this QTMS index was procedded by using nine bonds. To obtain QSPR equation for each QTMS descriptor, MLR analysis with stepwise selection of variables was used. The results are listed in Table 4. As is seen, significant correlation models have been obtained for all QTMS indices except $\bigtriangledown^2\rho$ descriptor. Therefore, it can be concluded that all calculated QTMS indices, except $\bigtriangledown^2\rho$, are useful parameters in describing the O-methylation kinetics of the phenol derivatives. The resulted models have high prediction ability and generalizations as obtained by cross-validation (R^2_{CV}) and using a separate test set (R^2_P). Column 3 of Table 4 shows for each QTMS the chemicals bonds whose corresponding QTMS parameter has shown significant effect on the rate constant. As an example in model number 5 of Table 4, the λ_3 of C1—O7 and O7—H8 are affected the rate constant much higher than the λ_3 of the other bonds. It is observed that C1—O7 and O7—H8 or combination of them have appeared in almost all of the models listed in Table 4. These findings are in direct agreement with those obtained in previous section (Table 3).

In the third approach for analyzing the QTMS descriptors, both bond information and descriptor information were used to generate the QSPR models. For this purpose, the matrix of the calculated descriptors of each molecule was unfolded to a row vector by collecting the QTMS descriptors of each bond beside each others. In this case, all calculated descriptors are used in the QSPR model development. Since the number of calculated descriptors (76 descriptors after removing the collinear variables) is much higher than the number of molecules (36 molecules), PLS regression was employed instead of MLR analysis. It should be noted that removing collinear descriptors is not essential in PLS analysis, however, it reduces dataset dimension and decreases the redundancy existed in the descriptors data matrix. The model refinement procedure was used LOO-CV to select the optimum number of PLS latent variables. Once the optimum number of latent variables was determined and the PLS model was built, the data are randomly classified into calibration (26 molecules) and prediction (10 molecules) sets. The PLS regression coefficients at optimum number of factors were calculated using the calibration samples and then were used to calculate the $\log(k_X/k_H)$ of the prediction set molecules. At the first trial, no feature selection method was employed and therefore all 76 parameters were used in model development. The statistical parameters for the resulted PLS models are summarized in Table 5. As it is observed, this model has high calibration statistical quality (i.e. $R^2 = 0.959$) whereas it has low prediction ability measured by cross-validation and using a separate test set (i.e. R^2_{CV} and R^2_p are equal to 0.665 and 0.544, respectively). A large difference between the prediction and calibration statistics (high calibration and low prediction statistics) indicates the presence of overfitting problem in the resulted model.[37] Two main sources of overfitting in the factor analysis-based regression method are the use of higher number of latent variables and higher number of predictor variables than they are necessary.[37] Here, the number of latent variables was optimized by cross-validation procedure, which currently is a standard method for optimizing model complexity. The use of calibration statistics in optimizing the PLS latent variables generally leads to overfitted model.[37] As an example the variation of the prediction residual errors sum of squares (PRESS) from calibration data or from cross-validation as a function of the PLS latent variables has been plotted in Figure 2. The plots show the gradual decreasing of PRESS of calibration while number of latent variables is increased. In the other hand the plot of PRESS of cross-validation as a function of latent variables is passed through a distinct minimum at number of latent variables of nine.

According to the above results, it can be concluded that the major source of overfitting in the PLS model mentioned above

Table 5. Statistical Parameters for Different PLS Models.

Method	f	R^2	R^2_{CV}	R^2_P
PLS	12	0.959	0.665	0.544
GA-PLS	7	0.981	0.970	0.961
PC_{Bond}-PLS	4	0.927	0.918	0.924
PC_{QTMS}-PLS	3	0.942	0.937	0.931

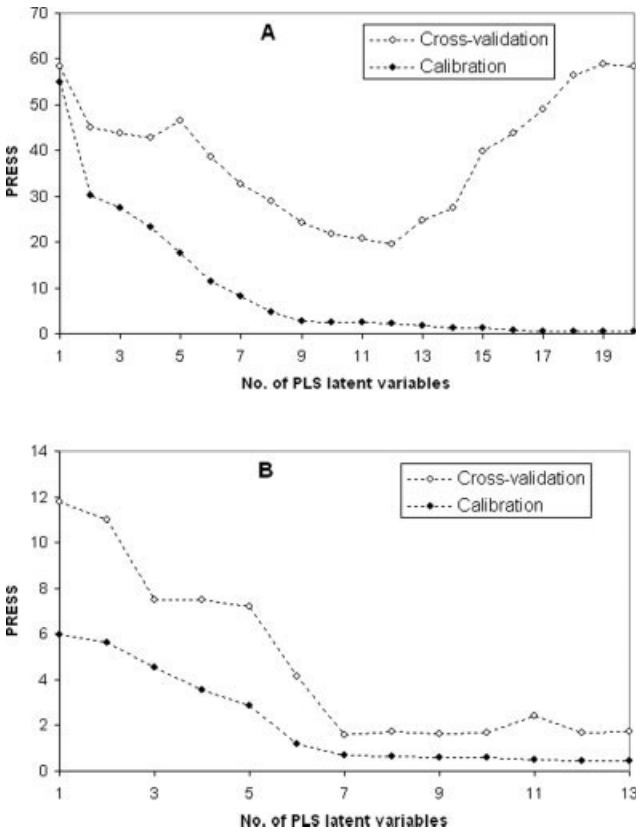

Figure 2. PRESS plot for PLS (A) and GA-PLS (B) models.

is using higher numbers of predictor variables than they are necessary. According to the results obtained in the first and second approaches of data analysis, only some of descriptors (among the 76 calculated descriptors) have significant effects on the rate constant of O-methylation. Detecting these variables will help us to obtain a more deep insight about the significance of QTSM properties and molecular bonds in the chemical process under study. Currently, some variable selection methods are available that can select the most relevant set of descriptors among a large number of calculated descriptors.[38–41] Among them, GA has found the major popularity in the chemical literature.[34,35,42–44] A GA is a problem solving method that uses genetic rules such as reproduction, crossover and mutation to build pseudo organisms that are then selected, on the basis of a fitness criterion to survive and pass information on to the next generation.[42–44] The statistical quantity of the most convenient GA-PLS model that produced the best cross-validation and prediction statistics is represented in Table 5. GA has selected 17 variables for PLS to model the linear relationships between $\log(k_X/k_H)$ and the QTSM descriptors. The number of latent variables is seven, which is not significantly different from the previous case where all descriptors were used. The plots of PRESS have been represented in Figure 2 and again the usefulness of cross-validation in determining the optimum number of latent variables is confirmed. It is clearly observed from Table 5 that GA-PLS has similar calibration statistics with respect to the conventional PLS; however, the former has improved cross-validation and prediction statistics. The squared correlation coefficients for cross-validation and prediction are 0.970 and 0.961, which indicates the ability of the model to reproduce more than 96% of variances in rate constant data.

To measure the significance of the 17 selected QSTM descriptors in the kinetic of the O-methylation, variable important in projection (VIP) was calculated for each descriptor.[45] VIP values reflect the importance of terms in the model. According to Erikson et al., X-variables (predictor variables) could be classified according to their relevance in explaining y (predicted variable) so that VIP > 1.0 and VIP < 0.8 mean highly or less influential, respectively, and $0.8 <$ VIP < 1.0 means moderately influential.[46] The calculated VIP values for the descriptors in the GA-PLS model are plotted in Figure 3. As is observed, 10 descriptors have VIP greater than one, which can be considered as highly effective parameters in the kinetic of the O-methylation of phenol derivatives. These parameters, in the decreasing order of VIP, are $K(r)$ (O7—H), $G(r)$ (C2—X), ε (C1—O7), λ_3 (C2—X), ρ (C1—C2), $K(r)$ (C4—X), ε (C5—C6), ε (O7—H), λ_3 (C2—C3), and λ_2 (O7—H). Obviously, the parameters are related to a restricted number of chemical bonds in the basic molecular skeleton of phenol (see Fig. 1B). Among them, C1—O7, O7—H, and C1—C2 are frontier bonds mostly involved in the O-methylation reaction whereas C2—X and C4—X, those of *ortho* and *para* substituents, respectively, have the highest electronic effects. For O7—H bond, three QTMS descriptors, including $K(r)$, ε, and λ_2, have been selected as highly effective whereas for C1—O7 only ε has been selected as the highly influential QTMS index. For the bond of para substituents (i.e., C4—X) Hamiltonian kinetic energy ($K(r)$) is selected. Those descriptors for C2—X are $G(r)$ and λ_3 and that of C2—C1 is ρ.

The PLS-regression coefficients for the normalized variables are represented in Figure 4. It shows that how do the selected QTMS affect the reaction kinetic. As it is observed, among the 17 selected descriptors, 6 variables including $G(r)$ (C2—X), $K(r)$ (C4—X), λ_2 (O7—H), ρ (O7—H), ε (C1—C2), and ε (O7—H) have negative coefficient and the rest represent positive effect. Smith and Popelier[47] related the changing in QTMS indices to the Hammett's electronic substituent constant (σ). Therefore, the dependency of O-methylation rate constant of the substituted phenols on the some selected QTMS indices can be attributed to the correlation between the substituent constants and the QTMS indices. For example, for the four-substituted phenols, we obtained a reverse relationship (with $R^2 = 0.83$) between $K(r)$ (O7—H) and Hammet's σ constant. On the other hand, electron-withdrawing substituents on the phenol ring can stabilize the phenoxide anion and as it is observed from Table 1, decreased rate constants have been determined for the molecules with such substituents, having high value of σ constant. Therefore, a negative effect of σ substituent constant on the reaction rate is expected. Accordingly, the positive effect of $K(r)$ (O7—H) on the reaction kinetic can be attributed the negative correlations between this QTMS index and σ constant in one hand and reaction rate and σ constant in the other hand. Similar discussion can be given for the negative effect of $G(r)$ (C2—X) and $K(r)$ (C4—X) as these descriptors represent direct relationship with σ substituent constant.

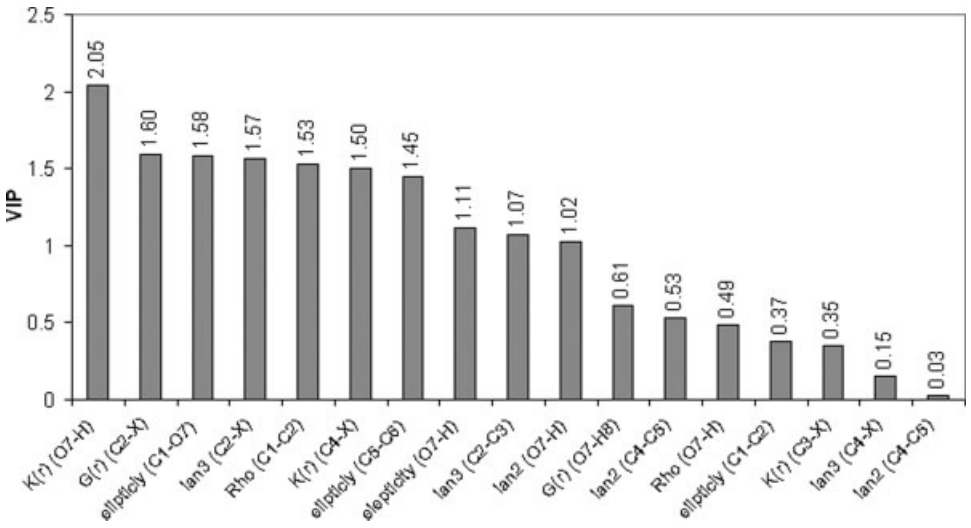

Figure 3. Plot of variable important in projection (VIP) for the QTMS indices of the chemical bonds obtained by GA-PLS model.

To handle the QTMS descriptor data matrix of a single molecule, Popelier and coworkers have already used principal components analysis (PCA), compressing the descriptor data matrix of a single molecule into an eigenvector named principal component.[20–26] For a given data matrix, PCA produces two set of eigenvectors named row and column vectors,[33] which are usually named score and loading vectors. These vectors span the row and column spaces of the original data matrix, respectively. In the case of the QTMS descriptor data matrix of a molecule and according to the data arrangement used in this article, the row and column vectors span the bond and QTMS descriptors spaces, respectively. These eigenvectors are the basis of predictive and descriptive QSPR models. Popelier and coworkers used row vectors to account the effect of chemical bonds on the property or activity of interest.[20–26] In this work, we used both the row and column vectors to account the bond and descriptors effect separately. These are the basis of the fourth and fifth approaches in the analysis of QTMS data. The main difference between these approaches and the third one (previous paragraph) is that the row and column vectors contain information solely from bonds and QTMS descriptors, respectively. This allows us to separately investigate the effect of bonds and descriptors.

The results of the application of PCA on the QTMS descriptors data matrix of some representative phenol derivatives are listed in Table 6. In this table, the eigenvalues, the percent of variances in the descriptors data matrix explained by each eigen-

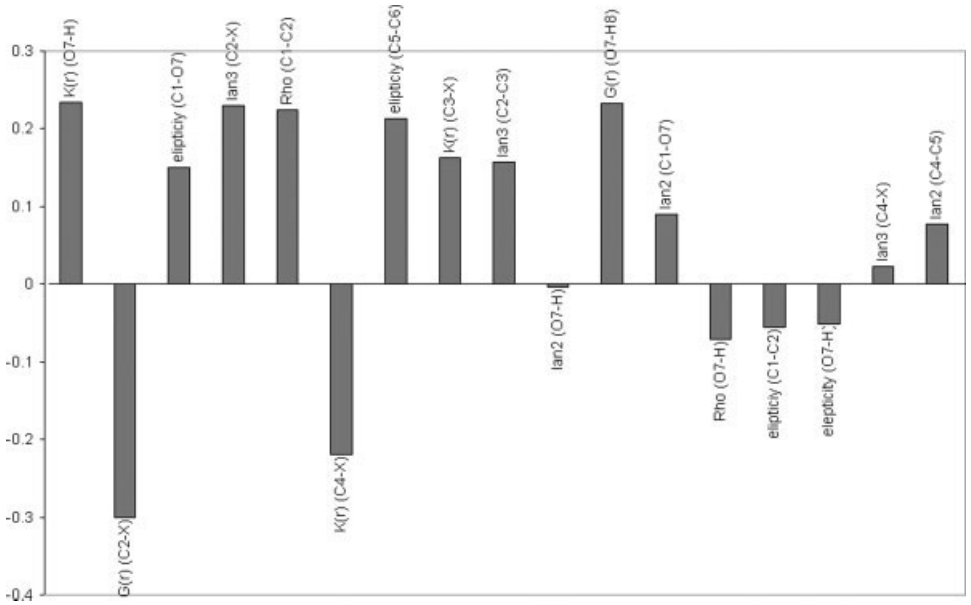

Figure 4. Plot of the PLS regression coefficients for the descriptors selected by GA-PLS.

Table 6. Results of Application of PCA on the QTMS Descriptor Data Matrices of Some Representative Phenol Derivatives.

	Eigenvalue				Percent of variance				Cumulative percent of variance			
Derivative	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4
Phenol	5.15	1.59	1.26	0.004	64.3	19.8	15.8	0.05	64.3	84.2	99.9	99.99
2-$CONH_2$	5.22	1.50	1.23	0.05	65.2	18.8	15.3	0.62	65.2	84.0	99.3	99.95
2-Me	5.17	1.56	1.26	0.02	64.6	19.5	15.7	0.19	64.6	84.0	99.8	99.95
3-F	4.63	2.22	1.07	0.08	57.8	27.8	13.4	0.97	57.8	85.6	99.0	99.99
3-OMe	4.94	1.96	1.10	0.01	61.7	24.4	13.8	0.07	61.7	86.2	99.9	99.99
4-OH	4.96	1.93	1.11	0.00	62.0	24.1	13.9	0.02	62.0	86.1	99.9	99.98
4-NO_2	5.17	1.46	1.27	0.11	64.6	18.2	15.9	1.32	64.6	82.8	98.7	99.98
2-OMe-4-Me	4.97	1.84	1.17	0.01	62.1	22.9	14.7	0.15	62.1	85.1	99.7	99.88
2,4-NO_2	5.22	1.47	1.12	0.17	65.3	18.4	14.0	2.10	65.3	83.7	97.7	99.79

value, and the cumulative percent of variances have been listed for the first to fourth principal components. As is observed, the largest variances have been explained by the first PC (about 60%) and the second and third PCs explain lower and relatively similar variances (between 13 and 27%). The forth PC covers very low variances in the QTMS descriptors data matrix. The first three PCs can explain more than 99% of variances in the data matrices and there is a large difference between the eigenvalue of the third and fourth PCs. These indicate the presence of three significant sources of variances in the QTMS descriptor data matrix for each molecule. However, for the sake of simplicity, only the first PC or eigenvector was used to derive QSPR models. PCA-based PLS regression was performed twice using the first row eigenvector as predictor variable in one hand and the first column eigenvector on the other hand.

As noted previously, the row eigenvectors contain QTMS information about the chemical bonds, and thus each eigenvector has a size of 13. The data matrix of independent variables therefore has a dimension of (36 × 13). Modeling of the relationship between the principal components of bonds and rate constant was performed by PLS regression using entire set of the principal components of the chemical bonds. The statistical parameters for the resulted QSPR model are listed in Table 5. The high correlation coefficients of cross-validation and prediction explain the high prediction ability of the resulted model. However, the model performance is moderately lower than that obtained by the use of original descriptors as predictor variables (GA-PLS model). The plot of VIP for the principal components of the bonds is shown in Figure 5. It is observed from this plot that the order of the importance of the bonds in the O-methylation rate constants of the phenol derivatives is O7—H8 > C1—O7 > C2—X > C4—X > C1—C2 ≈ C1—C6. The VIP value of the other bonds is lower than 0.8, and so they can be considered as noninfluential parameters. These bonds are found to be significant by GA-PLS method (Fig. 1B).

In another trial, the first column vector of the QTMS descriptors data matrix of each molecule was used to investigate the sole effect of the calculated QTMS descriptors without using bond information. A multilinear QSAR model between the principal components of the descriptors (PC_{QTMS}) and rate constant was obtained by PLS regression. The statistical quality of the resulted model (PC_{QTMC}-PLS), summarized in Table 5, confirms the high predictivity of this model. As is observed, the quality of the PC_{QTMS}-PLS model is slightly better than that of PC_{Bond}-PLS model. This explains the importance of QTMS indices in the modeling of O-methylation rate constant of substituted phenols. The VIP plot for the principal components of the QTMS indices (Fig. 6) shows that all QTMS indices can be considered as influential parameters whereas among them, λ_1, λ_2, λ_3, ε, and $K(r)$ are the most significant ones, and $\bigtriangledown^2\rho$, ρ, and $G(r)$ have moderate significance.

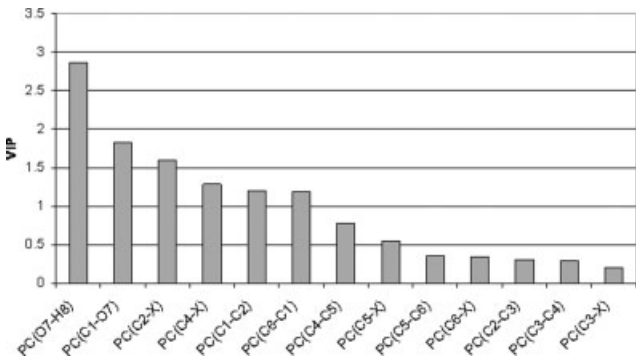

Figure 5. Plot of variable important in projection (VIP) for the principal components of the bonds obtained by PC_{Bond}-PLS model.

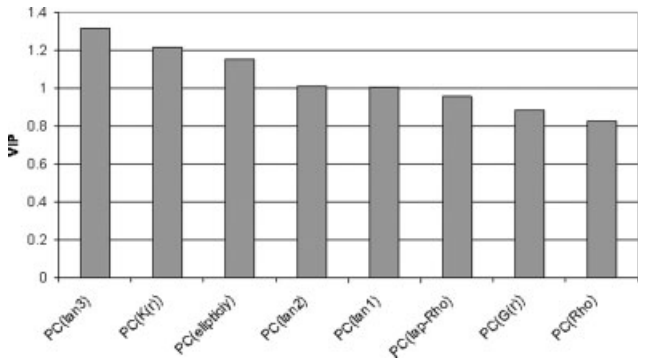

Figure 6. Plot of variable important in projection (VIP) for the principal components of the QTMS indices obtained by PC_{QTMS}-PLS model.

Conclusion

A set of novel quantum chemical descriptors called QTMS indices was used to derive QSPR models for the kinetic of the O-methylation reaction of substituted phenols. Some different chemometrics methods including PCA, PLS, and GA were used in the modeling processes. The resulted models could reproduce about 95% of variances in the rat constant data. QTMS theory produces a matrix of descriptors containing both bond (structure) and electronic information. By the use of PCA, postprocessed by PLS and then calculating the VIP for each parameter, the effect of these two distinct features on the rate constant was investigated. The results revealed that the rate constant of esterification of phenols is highly influenced by the electronic properties of the C2—C1—O—H fragment of the parent molecule, which can be considered as frontier bonds in the O-methylation reaction (Fig. 1B). Indeed, the C2—X and C4—X bonds (corresponding to ortho and para substituents) were found as highly influential parameters. All of the eight calculated QTMS indices were found significant however, λ_1, λ_2, λ_3, ε, and $K(r)$ were detected as highly influential parameters.

References

1. Bultinck, P.; De Winter, H.; Langenaeker, W.; Tollenaere, J. P. Computational Medicinal Chemistry for Drug Design; Marcel Decker: New York, 2004.
2. Hansch, C.; Leo, A. Exploring QSAR: Fundamentals and Applications in Chemistry and Biology; ACS: Washington, DC, 1995.
3. Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. Nature 1962, 194, 178.
4. Jalali-Heravi, M.; Kyani, A. J Chem Inf Comput Sci 2004, 44, 1328.
5. Hemmateenejad, B.; Miri, R.; Safarpour, M. A.; Mehdipour A. R. J Comput Chem 2006, 27, 1125.
6. Habibi-Yangjed, A.; Danandeh-Jahangard, M.; Nooshyar, M. J Mol Model 2006, 12, 338.
7. Xi, Z.; Yu, Z.; Niu, C.; Ban, S.; Yang, G. J Comput Chem 2006, 27, 1571.
8. Todeschini, R.; Consoni, V. Handbook of Molecular Descriptors; Wiley-VCH: Weinheim, 2000.
9. Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Chem Rev 1996, 96, 1027.
10. Parthasarathi, R.; Padmanabhan, J.; Subramanian, V.; Maiti, B.; Chattaraj, P. K. J Phys Chem A 2003, 107, 10346.
11. Turrabekova, M. A.; Rasulev, B. F. Molecules 2004, 9, 1194.
12. Platts, J. A. Phys Chem Chem Phys 2000, 2, 3115.
13. Zhang, S. G.; Lei, W.; Xia, M. Z.; Wang, F. Y. J Mol Struct (Theochem) 2005, 732, 173.
14. Hemmateenejad, B.; Safarpour, M. A.; Miri, R.; Taghavi, F. J Comput Chem 2004, 25, 1495.
15. Lahsen, J.; Schmidhammer, H.; Spetea, M.; Rode, B. M. QSAR Comb Sci 2003, 22, 476.
16. Takashima, H.; Kitamura, K.; Tanabe, K.; Nagashima, U. J Comput Chem 1999, 20, 443.
17. Popelier, P. L. A. J Phys Chem A 1999, 103, 2883.
18. O'Brein, S. E.; Popelier, P. L. A. J Chem Inf Comput Sci 2001, 41, 764.
19. Bader, R. F. W. Atoms in Molecules: A Quantum Theory; Oxford University Press: Oxford, 1990.
20. Popelier, P. L. A. Atom in Molecules. An introduction; Pearson Education: London, 2000.
21. O'Brein, S. E.; Popelier, P. L. A. J Chem Soc Perkin Trans 2002, 2, 478.
22. Chaudry, U. A.; Popelier, P. L. A. J Phys Chem A 2003, 107, 4578.
23. Chaudry, U. A.; Popelier, P. L. A. J Org Chem 2004, 69, 233.
24. Popelier, P. L. A.; Chaudry, U. A.; Smith, P. J. J Chem Soc Perkin Trans 2002, 2, 1231.
25. Loader, R. J.; Singh, N.; O'Malley, P. J.; Popelier, P. L. A. Bioorg Med Chem Lett 2006, 16, 1249.
26. Popelier, P. L. A.; Smith, P. J. Eur J Med Chem 2006, 41, 862.
27. Selassie, C. D.; Garg, R.; Kapur, S.; Kurup, A.; Verma, R. P.; Mekapati, S. B.; Hansch, C. Chem Rev 2002, 102, 2585.
28. Ren, S.; Kim, H. J Chem Inf Comput Sci 2003, 43, 2106.
29. Garg, R.; Kapur, S.; Hansch, C. Med Res Rev 2001, 21, 73.
30. Cork, D. G.; Hayashi, N. Bull Chem Soc Jpn 1993, 66, 1583.
31. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, W. M.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. Gaussian 98, Revision A.7; Gaussian, Inc.: Pittsburgh, PA, 1998.
32. AIM2000, Version 2.0, 2002.
33. Malinowski, E. R. Factor Analysis in Chemistry; Wiley: New York, 2002.
34. Hemmateenejad, B. Akhond, M.; Miri, R.; Shamsipur, M. J Chem Inf Comput Sci 2003, 43, 1328.
35. Hemmateenejad, B. J Chemometr 2004, 18, 475.
36. Fernandez, M.; Caballero, J. J Mol Graph Model 2006, 25, 410.
37. Hawkins, D. M. J Chem Inf Comput Sci 2004, 44, 1.
38. Chong, I.; Jun, C. H. Chemometr Intell Lab Syst 2005, 78, 103.
39. Livingstone, D. J.; Salt, D. W. Rev Comput Chem 2005, 21, 287.
40. Liu, Y. J Chem Inf Comput Sci 2004, 44, 1823.
41. Shamsipur, M.; Zareh-Shahabadi, V.; Hemmateenejad, B.; Akhond, M. J Chemometr 2006, 20, 146.
42. Hou, T.; Xu, X. Prog Chem 2004, 16, 35.
43. Leardi, R. J Chemometr 2001, 15, 559.
44. Hibbert, D. B. Chemometr Intell Lab Syst 1993, 19, 277.
45. Olah, M.; Bologa, C.; Opera, T. I. J Comput Aided Mol Des 2004, 18, 437.
46. Erikson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S.; Multi- and Megavariate Data Analysis. Principle and Applications; Umetrics Academy: Umea, 2001.
47. Smith, P. J.; Popelier, P. L. A. Org Biomol Chem 2005, 3, 3399.