Journal of Computer-Aided Molecular Design (2005) 19: 693–703
DOI 10.1007/s10822-005-9008-0

Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures

A. Varnek[a,]*, D. Fourches[a], F. Hoonakker[a] & V. P. Solov'ev[b]
[a]*Laboratoire d'Infochimie, UMR 7551 CNRS, Université Louis Pasteur, 4, rue B. 67000, Pascal, Strasbourg, France;* [b]*Institute of Physical Chemistry, Russian Academy of Sciences, Leninskiy Prosp. 31a, 119991 Moscow, Russia*

Received 17 May 2005; accepted 28 July 2005

Key words: chemical reactions, clustering, condensed graphs of reactions, exchange of information, hydrogen bonds, QSPR, similarity, solubility, substructural fragments

Summary

Substructural fragments are proposed as a simple and safe way to encode molecular structures in a matrix containing the occurrence of fragments of a given type. The knowledge retrieved from QSPR modelling can also be stored in that matrix in addition to the information about fragments. Complex supramolecular systems (using special bond types) and chemical reactions (represented as Condensed Graphs of Reactions, CGR) can be treated similarly. The efficiency of fragments as descriptors has been demonstrated in QSPR studies of aqueous solubility for a diverse set of organic compounds as well as in the analysis of thermodynamic parameters for hydrogen-bonding in some supramolecular complexes. It has also been shown that CGR may be an interesting opportunity to perform similarity searches for chemical reactions. The relationship between the density of information in descriptors/knowledge matrices and the robustness of QSPR models is discussed.

Introduction

Nowadays, a safe exchange of data associated with chemical compounds but without revealing their structures is very desirable because it provides the academic or industrial researchers with the data required for structure-property studies, whereas the owner of the data obtains the results of research without any risk to lose the confidentiality of information.

Fragment descriptors obtained from 2D graphs provide a possibility to encode chemical structures [1–3] as well as to perform structure – property studies being used as variables in a multi-linear regression [1, 4–14] or in neural networks [8, 10]. Computation of fragments does not require knowledge of the geometry and electronic structure of molecules, and structural fragments are more easily interpretable than topological indices or some physico-chemical descriptors. Many different types of fragment descriptors have been suggested: the sequences of atoms and bonds; "augmented" atoms including a "central" atom with its several topological coordination spheres; branched fragments and small rings. The "molecular signature" descriptors of Faulon [1, 15, 16], representing a combination of augmented atoms and atom/bond sequences, have been successfully used both to encode chemical structures and to reconstruct them from the set of descriptors. Substructural fragments are closely related to topological descriptors; the latter can be represented as a linear combination of occurrences of some substructures [17, 18].

*To whom correspondence should be addressed. E-mail: varnek@chimie.u-strasbg.fr

An important requirement for sharing information is the pertinence of molecular descriptors which map structural information with properties. In this sense, substructural fragments are good candidates. They have been successfully used in diversity analysis of large databases [4, 19] and in structure-property studies [1, 4, 20–25]. It should be noted that QSAR methods based on 2D fragment descriptors represent an appealing alternative to 3D QSAR since they do not require extensive conformational analysis and spatial alignment of molecules. They are faster and easier to implement in an automated fashion and are typically characterized by the same or better statistics compared to 3D QSAR methods [26, 27].

Here we demonstrate that substructural fragments can be used as descriptors not only for a set of individual compounds, but also for more complex supramolecular structures as well as for chemical reactions. The paper contains two parts. The first one describes the general approach of encoding structures and reactions using substructural fragments, and gives some information about particular fragments and developed software tools. The second part is devoted to application of fragment descriptors to assess physico-chemical properties of molecules (aqueous solubility) and supramolecular ensembles (thermodynamics of hydrogen bond complexes) as well as to perform a similarity search for reactions.

Encoding reactions, molecular and supramolecular structures using substructural fragments

Under the fragment approach, a molecular structure can be encoded by the vector constituted from occurrences of the fragments of each type. Consequently, the set of molecules is encoded by a descriptors (pattern) matrix which combines the vectors encoding individual molecules (Figure 1). When the property values are added, the resulting *descriptors/properties* matrix characterizes both structures and properties of the compounds in the given data set. If substructural fragments are used as descriptors in QSPR studies, the modelled property can be presented as a linear combination of selected descriptors. Thus, the *descriptors/knowledge* matrix (Figure 1) contains information only about those descriptors which are pertinent to the modelled property. Thus, from the same descriptors/properties matrix, one can obtain

$$\left(\begin{array}{c|cccc} & D_1 & D_2 & \cdots & D_k \\ \hline Mol_1 & N_{11} & N_{12} & \cdots & N_{1k} \\ Mol_2 & N_{21} & N_{22} & \cdots & N_{2k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ Mol_n & N_{n1} & N_{n2} & \ldots & N_{nk} \end{array}\right)$$

descriptors matrix

$$\left(\begin{array}{c|ccccccccc} & D_1 & D_2 & \cdots & D_k & \vdots & PR_1 & PR_2 & \cdots \\ \hline Mol_1 & N_{11} & N_{12} & \cdots & N_{1k} & \vdots & P_{11} & P_{12} & \cdots \\ Mol_2 & N_{21} & N_{22} & \cdots & N_{2k} & \vdots & P_{21} & P_{22} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \vdots & \cdots & \cdots & \cdots \\ Mol_n & N_{n1} & N_{n2} & \ldots & N_{nk} & \vdots & P_{n1} & P_{n2} & \cdots \end{array}\right)$$

descriptors /properties matrix

$$\left(\begin{array}{c|ccccc} & \cdots & D_j & \cdots & D_r \\ \hline Mol_1 & \cdots & N_{1j} & \cdots & N_{1r} \\ Mol_2 & \cdots & N_{2j} & \cdots & N_{2r} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ Mol_n & \cdots & N_{nj} & \cdots & N_{nr} \\ \hline Koef & \cdots & K_j & \cdots & K_r \end{array}\right)$$

descriptors /knowledge matrix

Figure 1. Encoding molecular structures and properties using fragment descriptors. Here, N_{ij} is the number of fragments of j-type (D_j) in the molecule i (Mol$_i$), P_{ij} is the value of the property j (PR_j) for i-th molecule, K_j is the coefficient at D_j in the multi-linear equation (1), $r \le k$.

several different descriptors/knowledge matrices, each of them being related to a given property.

If the owner of the data does not wish to reveal the structures, information about types of constituent fragments in the descriptors matrix can be hidden. Reconstruction of the structures from the descriptors matrix or descriptors/property matrix then becomes hardly possible (Figure 2). On the other hand, the researcher is able to run QSPR studies producing the descriptors/knowledge matrix. Thus, the confidentiality of the structural information does not prevent the researcher extracting knowledge from the owner's data.

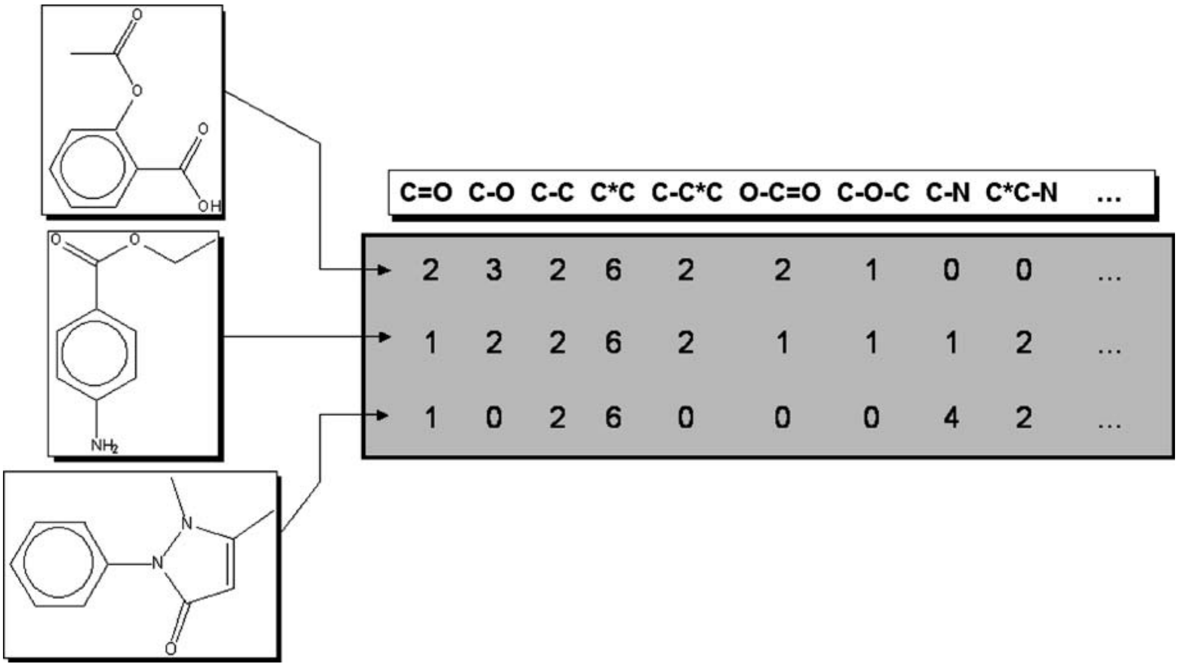
In our previous publications [5, 7, 9, 11–13], we suggested the use in QSPR studies of two different classes of substructural molecular fragments: "sequences" (**I**) and "augmented atoms" (**II**). Three sub-types, **AB**, **A** and **B** are defined for each class. For the fragments **I**, they represent sequences of atoms and bonds (**AB**), of atoms only (**A**), or of bonds only (**B**). Only the shortest paths from one atom to the other are used. For each type of sequence, the minimum ($n_{\min}$) and maximum ($n_{\max}$) number of constituent atoms must be defined. Thus, for the partitioning **I**(**AB**, $n_{\min}$–$n_{\max}$), **I**(**A**, $n_{\min}$–$n_{\max}$) and **I**(**B**, $n_{\min}$–$n_{\max}$), the program generates "intermediate" sequences involving n atoms ($n_{\min} \leq n \leq n_{\max}$) (Figure 3).

An "augmented atom" represents a selected atom with its environment including either neighbouring atoms and bonds (**AB**), or atoms only (**A**), or bonds only (**B**). Atomic hybridization (**Hy**) can be taken into account for augmented atoms of the **A**-type (Figure 3). Supramolecular structures can be treated in a similar way, if the hydrogen bonds, coordination bonds or any other types of bonds are represented explicitly (see Table 1).

This fragment approach can be extended to chemical reactions using *Condensed Graphs of Reaction* (CGR), [29–32] in which reactants and products are "condensed" into one 2D graph involving both conventional and "dynamic" bonds (Figure 4). This provides users with an opportunity to treat an ensemble of reacting species as one pseudo-compound (see Section 4.3).

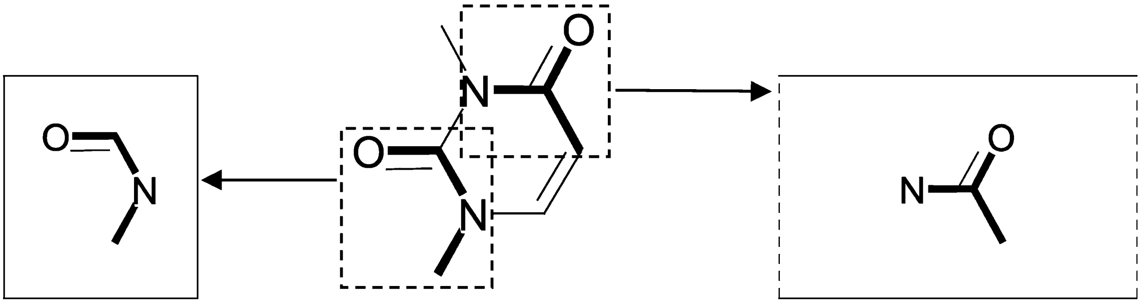
Tools for the mining of chemical data using fragment descriptors

The ISIDA (In Silico Design and Data Analysis) package has been developed to perform structure-



C=O	C-O	C-C	C*C	C-C*C	O-C=O	C-O-C	C-N	C*C-N	...
2	3	2	6	2	2	1	0	0	...
1	2	2	6	2	1	1	1	2	...
1	0	2	6	0	0	0	4	2	...

Figure 2. Example showing a part of the descriptors matrix for three organic compounds. The integer numbers correspond to occurrences of the constituent substructural fragments. If the list of these fragments is suppressed, the structure of molecules can hardly be revealed.



SEQUENCES I		AUGMENTED ATOMS II
	ATOMS and BONDS (AB)	
C-N-C=O; C-N-C; N-C=O; 2 C-N; C=O		C (-C) (-N) (=O)
	ATOMS (A)	
C N C O; C N C; N C O; 2 C N; C O	(Hy)	C (C) (N) (O) or $C_{sp2}(C_{sp3})(N_{sp3})(O_{sp2})$
	BONDS (B)	
- - =; - -; - =; 2 -; =		C (-) (-) (=)

Figure 3. Two classes of substructural fragments: atom/bond sequences and augmented atoms. Shortest paths sequences (**I**) and augmented atoms (**II**) including atoms and bonds (**AB**), only atoms (**A**) or only bonds (**B**). From top to bottom: the sequences (**I**) correspond to the **I**(**AB**, 2–4), **I**(**A**, 2–4) and **I**(**B**, 2–4) types involving paths between each pair of atoms. The **II**(**Hy**) augmented atoms correspond to the **II**(**AB**) type, where hybridization of atoms is taken into account.

property and clustering studies using fragment descriptors [33]. ISIDA includes QSPR, clustering and combinatorial modules as well as some supplementary tools including the editor of 2D structures, *EdChemS*, the editor of SD files, *EdiSDF*, and converters of formats, etc. For molecular and supramolecular structures, the *EdChemS* editor recognizes 10 different types of bonds: single, double, triple (in cycles or in chains), aromatic bonds and three types of coordination bonds (Table 1). Nine types of dynamic bonds related to transformations of one type of conventional bond to another one can be used for preparation of condensed graphs of reactions (Table 1, Figure 4).

Structure-property calculations. Once a molecular graph is split into constitutive fragments, any corresponding quantitative physical or chemical property **PR** is calculated from the fragments contributions using linear (1) or non-linear (2) and (3) fitting equations:

$$\mathbf{PR} = a_0 + \sum_i a_i N_i + \Gamma \tag{1}$$

$$\mathbf{PR} = a_0 + \sum_i a_i N_i + \sum_i b_i(2N_i^2 - 1) + \Gamma \tag{2}$$

$$\mathbf{PR} = a_0 + \sum_i a_i N_i + \sum_{i,k} b_{ik} N_i N_k + \Gamma \tag{3}$$

Table 1. Bond types for compounds, complexes and reactions implemented in the ISIDA package.

	No.	Bond code[a]	Comments
Compounds	1	1	A single bond in chain
	2	2	A double bond in chain
	3	3	A triple bond in chain
	4	4	An aromatic bond
	5	5	A single bond in cycle
	6	6	A double bond in cycle
	7	7	A triple bond in cycle
Complexes	8	8	Coordination bond, type I
	9	9	Coordination bond, type II
	10	10	Coordination bond, type III
Reactions	11	81	Transformation of “no bond” to single bond
	12	12	Transformation of single bond to double bond
	13	23	Transformation of double bond to triple bond
	14	13	Transformation of single bond to triple bond
	15	18	Transformation of single bond to “no bond”
	16	21	Transformation of double bond to single bond
	17	28	Transformation of double bond to “no bond”
	18	32	Transformation of triple bond to double bond
	19	31	Transformation of triple bond to simple bond

[a]Bonds code in the connection table for MOL format [28].

where, a_i and b_i (b_{ik}) are fragment contributions, N_i is the number of fragments of type i. The a_o term is fragment independent. The a_i and b_i (b_{ik}) are the same for corresponding fragment for all compounds from the given set. An extra term $\Gamma = \Sigma c_m \text{ED}_m$ can be used to describe any specific feature of the compound using external descriptors ED_m (*e.g.*, topological, electronic, etc.); by default $\Gamma = 0$. Equation (1) represents a molecular property as a linear combination of fragment contributions. Equation (2), representing the three first terms of Chebyshev polynomial [34], accounts for non-additive effects related to individual fragments, whereas equation (3) involves a cross term $N_i N_k$ which accounts for the non-additivity effects of two different fragments. When non-linear equations (2) or (3) are used, the descriptors/knowledge matrix (Figure 1) contains columns corresponding not only to descriptors D_i, but also to terms ($2D_i^2$ - 1) or D_iD_i.

In the current version of ISIDA, the minimal and maximal lengths of the sequences are, respectively, $n_{\min} \geq 2$ and $n_{\max} \leq 15$. The number of types of sequences of different lengths for the range $n_{\min} = 2$ to $n_{\max} = 15$, is equal to 105 for each of three sub-types **AB**, **A** and **B**.

At the training stage, ISIDA builds up about 1300 structure-property models involving 3 linear and non-linear fitting equations and 319 types of fragment descriptors (*batch* calculations). If some fragments are linearly dependent, they are treated as one extended fragment. Using the singular value decomposition method (SVD) [35], ISIDA fits the a_i and b_i terms in equations (1) – (3) and performs statistical tests [36] to select the best models.

If some of the variables in equations (1)–(3) are linearly dependent or if a given fragment occurs in a relatively small number of molecules, the standard deviation Δa_i (Δb_i) for the fragment contributions a_i (b_i) can be large enough to lead to the corresponding t-test ($t = a_i/\Delta a_i$) being smaller than the tabulated value (t_0). The following procedure is applied in order to improve the robustness of the models. First, the program selects the variable with the smallest $t < t_0$, then it performs a new fitting excluding that variable. This procedure is repeated until $t \geq t_0$ for selected variables or if the number of variables attains the user’s defined value.

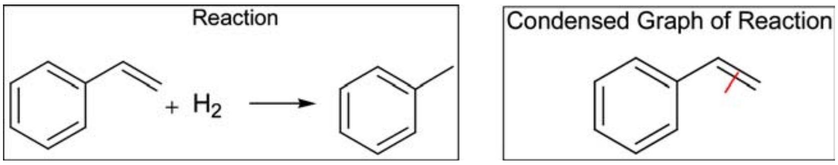


Figure 4. Encoding a reaction of hydrogenation into Condensed Graphs of Reaction. Dynamic bond “ ” corresponds to transformation of a double bond to a single bond (see Table 1).

Clustering. Partitioning of the structurally diverse data set into congener sub-sets followed by QSPR studies on each cluster may significantly improve the robustness of structure – property models. In this case, the calculations for one descriptors/properties matrix result in several descriptors/knowledge matrices. ISIDA performs clustering based on fragment descriptors using different metrics, linkages, normalizations and clustering algorithms. To treat large data sets, a successive use of the non-hierarchic Jarvis-Patrick algorithm and the hierarchical Johnson algorithm is suggested in order to reduce the time of calculations. Quite often, using conventional Euclidian or Manhattan distances (DIST_{ij}) between molecules *i* and *j* does not lead to chemically meaningful clusters. Therefore, modified distances (DIST_{ij}*) which enforce the fusion of similar compounds in one cluster, are suggested for the clustering, with fragment descriptors:

$$\mathrm{DIST}^{*}_{ij} = \mathrm{DIST}_{ij}(1 + 1/(T_{ij} + f)), \qquad (4)$$

where T_{ij} is a Tanimoto coefficient and $f = 0.05$ is an empirical fitting factor.

The content of clusters depends on the type of fragment descriptors. The choice of those fragments can be based on preliminary performed QSPR studies, thus resulting in clusters which fit the modelled property.

Structure-property and similarity studies using fragment descriptors

Assessment of aqueous solubility using combined clustering/QSPR approach

Aqueous solubility (log*S*) has been the subject of many QSPR studies [37–46] because of its major role in determining the bioavailability of compounds in organisms, and consequently, of its application in computer-aided drug design research. Here we show that combination of clustering and QSPR techniques based on fragment descriptors leads to the models which are, at least, as robust and predictive as the best models reported in the literature [37, 47].

A data set containing 1643 compounds was critically selected from the references [39, 48–50]. Clustering performed with ISIDA using the descriptors pool involving both **I**(**AB**,2–6) sequences and **II**(Hy) augmented atoms, resulted in four clusters containing 135 (A), 493 (B), 217 (C) and 798 (D) compounds (Figure 5, a). QSPR calculations were performed both on the entire set and on each cluster, resulting in several global and local models, respectively. The "global" I(AB, 2–5)/equation 1 model obtained in calculations on the full set has reasonable statistical parameters (R^2 = 0.924, RMSE = 0.60 and Q^2 = 0.900, Table 2). Calculations on clusters A, B and D resulted in the models based only on the sequences of atoms and bonds, whereas for the cluster C the models were based both on sequences and on augmented atoms (Table 2). The statistical parameters of these "local" models (R^2 = 0.803–0.952, RMSE = 0.36–0.59 and Q^2 = 0.618–0.922) were in some rare cases inferior to those obtained for the "global" model.

However, statistical parameters for linear correlations of log*S* (calc) vs log*S* (exp) of local models obtained for the clusters are clearly better than those of the global models obtained for the entire set (Figure 5, b, c). Thus, correlation coefficients R^2 obtained with local models are always higher than those obtained with the global model, whereas the opposite trend is observed for *RMSE* (Figure 5, b). Figures 5, c, d show that the ensemble of local assess log*S* for the entire set much better than the global model: the RMSE value for the linear correlation log*S* (calc) vs log*S* (exp) decreases from 0.60 (global) to 0.43 (local). The latter is similar to RMSE = 0.47 obtained for the set of 879 molecules using E-state indices and the neural networks technique of [51]. Thus, our multi-linear regression calculations on clusters resulted in models as efficient as those previously obtained with non-linear techniques.

Assessment of thermodynamics of intermolecular hydrogen bonds using labelled atoms

Assessment of thermodynamic parameters of the hydrogen-bond complexes is an important step toward a design of new supramolecular assemblies based on H-bond networks. In this section we demonstrate how substructural fragments can be used to assess the free energy (Δ*G*, kJ/mol) and enthalpy (Δ*H*, kJ/mol) of the 1:1 complexes between organic acids and bases linked by one hydrogen bond. Experimental data for 365

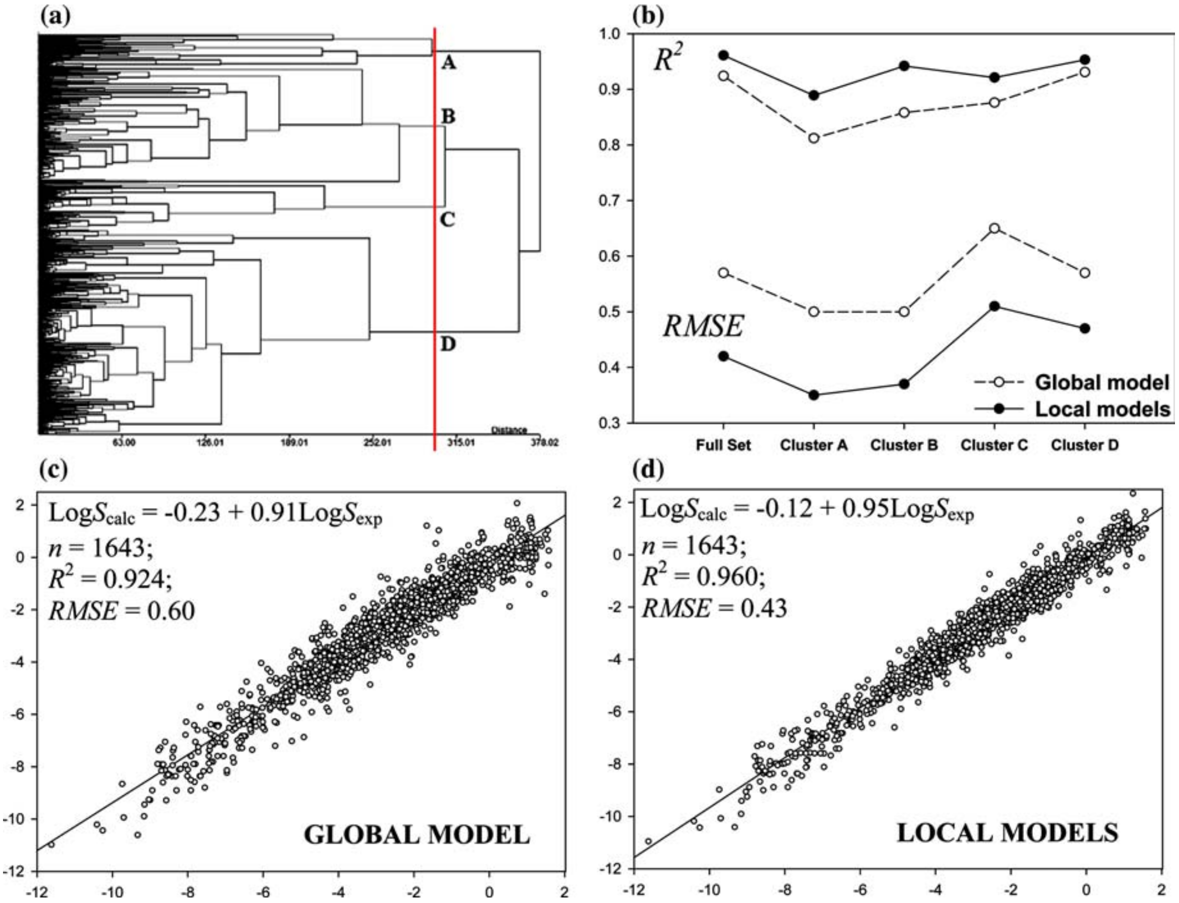


Figure 5. Studies of aqueous solubility (log*S*): (a) Clustering of the set of 1643 compounds performed using the descriptors pool involving both **I**(**AB**,2–6) sequences and **II**(Hy) augmented atoms, modified Euclidian distances, complete linkage and a combination of non-hierarchic (Jarvis-Patrick) and hierarchical (Johnson) algorithms. (b) Correlation coefficient R^2 and standard deviation *s* for linear correlations log*S* (calc) vs log*S* (exp) for the full set and clusters A, B, C and D. (c–d) Linear correlations log*S* (calc) vs log*S* (exp) obtained with QSPR models built either on the full data set (global models) or on clusters A, B, C and D (local models).

hydrogen bonded complexes in tetrachloromethane at 298 K were selected from the review by Raevsky et al. [52]. In these complexes, the acids are 46 substituted phenols, whereas the bases are represented by the large variety of chemical classes: phenols, alcohols, ethers, ketones, amides, heterocyclic compounds, phosporyl- and sulphonyl-containing molecules.

The *EdChemS* editor of 2D structures has been used to label all potential proton donor groups (OH groups in phenols) and proton acceptor groups (=O, –NH–, ...). Then, using the *EdiSDF* editor, an SD file containing structures of reagents and experimental values of ΔG and ΔH has been prepared and used in structure – property modelling. The hydrogen atoms were omitted in calculations.

The modelling of both free energy and enthalpy of the H-bond complexes resulted in the linear **I**(**AB**, 2–8)/equation 1 model based on the labelled atom-bond sequences containing from 2 to 8 atoms. In these calculations, the initial descriptors matrix containing 286 fragment descriptors was reduced to one containing 90 (ΔG) or 76 (ΔH) due to a variables selection procedure based on the *t*-test. These models have reasonable values of the correlation coefficient R^2 (fit) > 0.93, the *leave-one-out* cross-validation correlation coefficient Q^2 > 0.89 and RMSE = 0.70 kJ/mol for ΔG and 1.47 kJ/mol for ΔH (Table 3, Figure 6). Since

Table 2. Modelling of aqueous solubility (log*S*) for 1643 organic compounds: statistical criteria of selected linear models.[a,b]

Fragments	*N*	*k*	R^2 (fit)	RMSE (fit)	*F*	Q^2
			Full Set (1643 compounds)			
I(**AB**, 2–5)	1643	213	0.929	0.58	88.4	0.933
			Cluster A (135 compounds)			
I(**AB**, 4–4)	135	43	0.892	0.38	18.0	0.738
I(**AB**, 2–3)	135	30	0.803	0.52	14.8	0.618
			Cluster B (493 compounds)			
I(**AB**, 2–6)	493	92	0.946	0.36	77.9	0.884
			Cluster C (217 compounds)			
II(**Hy**)	217	43	0.906	0.58	40.0	0.858
I(**AB**, 2–4)	217	47	0.902	0.59	34.0	0.761
			Cluster D (798 compounds)			
I(**AB**, 2–5)	798	141	0.952	0.49	93.9	0.922

[a] Statistical parameters calculated for the training set: the number (*n*) of points (compounds), the number (*k*) of fitted coefficients in equation (1), correlation coefficient (*R*), root mean square error (RMSE), Fisher's criterion (*F*), factor of Hamilton (R_H), cross-validation correlation coefficient (*Q*).
[b] Molecules are represented without hydrogen atoms.

Table 3. Modelling of free energy (ΔG, kJ/mol) and enthalpy (ΔH, kJ/mol) for 365 H-bond complexes in CCl_4 at 298 K. Statistical criteria of the linear **I**(**AB**, 2–8)/equation 1 model.[a,b]

Property	Variables selection[c]	*k*	R^2	*F*	RMSE	Q^2
ΔG, kJ/mol	*1*	90	0.981	159.9	0.70	0.960
	2	50	0.950	122.3	1.13	0.936
	2	30	0.870	77.0	1.84	0.863
ΔH, kJ/mol	*1*	76	0.937	57.4	1.47	0.890
	2	50	0.876	45.5	2.07	0.843
	2	30	0.770	38.7	2.80	0.760

[a] See footnotes for Table 2.
[b] Fitting equation (1) was used with $a_0 = 0$. Only sequences containing labelled atoms were taken into account.
[c] The variables have been selected according *t*-test: (*1*) the calculations stop at $t > t_0$, and (*2*) the calculations stop at $t > t_0$ and at user's defined number of variables.

no coordination bonds between the acids and bases in the complexes were initially assumed, the resulting descriptors/knowledge matrix for the complexes is a superposition of two non-overlapping parts for the proton donors and acceptors, respectively.

In order to perform the internal validation of the model, the initial data set was split into the training set of 292 H-bond complexes and the test set of 73 complexes corresponding to each 5th complex from the initial set. Since some of compounds in the test set contained rare occurrence fragments, so that the validation calculations were not performed for those compounds. For the others, calculations with **I**(**AB**, 2–8)/equation 1 model show good correlations between "predicted" and experimental values: $\Delta G_{\mathrm{pred}} = 0.10 + 1.00\Delta G_{\mathrm{exp}}$ ($n = 66$, $R^2 = 0.916$, $F = 693$, $s = 1.67$) and $\Delta H_{\mathrm{pred}} = -2.28 + 0.90\Delta H_{\mathrm{exp}}$, ($n = 66$, $R^2 = 0.868$, $F = 421$, $s = 2.16$).

Thus, unlike as in additive-multiplicative models for enthalpy and stability constants of 1:1 hydrogen bond complexes previously developed by Drago [53, 54], Abraham [55–57] and Raevsky [58–61] we propose a simple additive scheme for calculations of ΔG and ΔH using contributions of selected fragments. This approach can be easily extended toward more complex systems involving multiple hydrogen bonds between organic acids and bases.

Density of information vs quality of QSPR models

A well-known problem of the fragment approach concerns the low density of information stored in the descriptors matrix. Indeed, for the set of 365 hydrogen bond complexes, the initial descriptors matrix built from 286 fragment descriptors contains $d = 4.6$ % of elements not equal to 0. The density of information increases with the decrease of the number of variables *k* selected in the QSPR models. Figure 7 shows that the reduction of *k* from 90 to 30 leads to an increase of *d* from 4.6 % to 13.4 %. On the other hand, such a reduction of *k* leads to less robust QSPR models. For example, in the modelling of ΔG, R^2 (fit) decreases from 0.981 to 0.870 and Q^2 decreases from 0.960 to 0.863 (Table 3), whereas the standard deviation for the linear correlation ΔG(calc) vs ΔG(exp) increases from 0.70 to 1.72 kJ/mol (Figure 6). For ΔH, this trend is even more pronounced (Table 3). Thus, one has to seek a compromise between the density of information in the descriptors matrix and related matrices (Figure 1) and the quality of QSPR models.

Similarity searching in the reactions space

Merging reactants and products of a chemical reaction into a Condensed Graph of Reaction

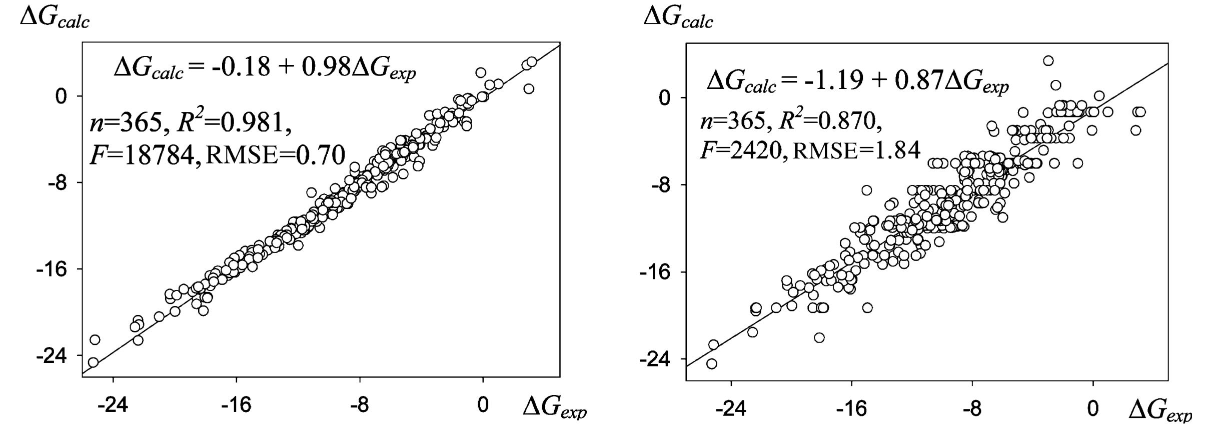


Figure 6. Modelling of free energy (ΔG, kJ/mol) for the set of 365 hydrogen-bond complexes in CCl_4 at 298 K. Linear correlation between calculated and experimental values for the **I**(**AB**, 2–8)/equation 1 model using 90 (left) and 30 (right) fragment descriptors selected by *t*-test.

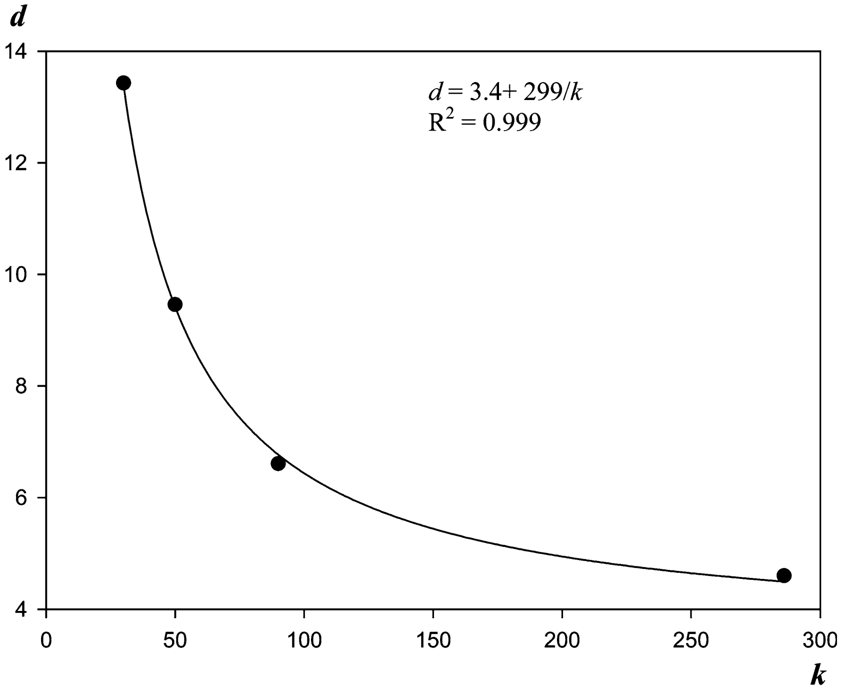


Figure 7. Modelling of thermodynamics parameters of the set of 365 hydrogen bond complexes: Density of information (d, %) in descriptors/knowledge matrix as a function of the number of fragment descriptors (k) selected according to *t*-test.

(CGR) gives the possibility of treating reactions as pseudo-compounds and, as a consequence, allows one to perform a similarity searching for queries represented as CGR. Recently, together with the Novalyst Discovery company [62] specializing in the development of new methods of organic synthesis, we have created a database containing several hundreds of hydrogenation reactions. Each record includes CGR, the reaction yield and experimental conditions (catalyst, temperature, pressure, solvent). Using ISIDA tools, one can search all CGRs having Tanimoto similarity coefficients larger than user's defined threshold (Figure 8). In such a way, for any hypothetical reaction, the user can search for the closest analogues in the database.

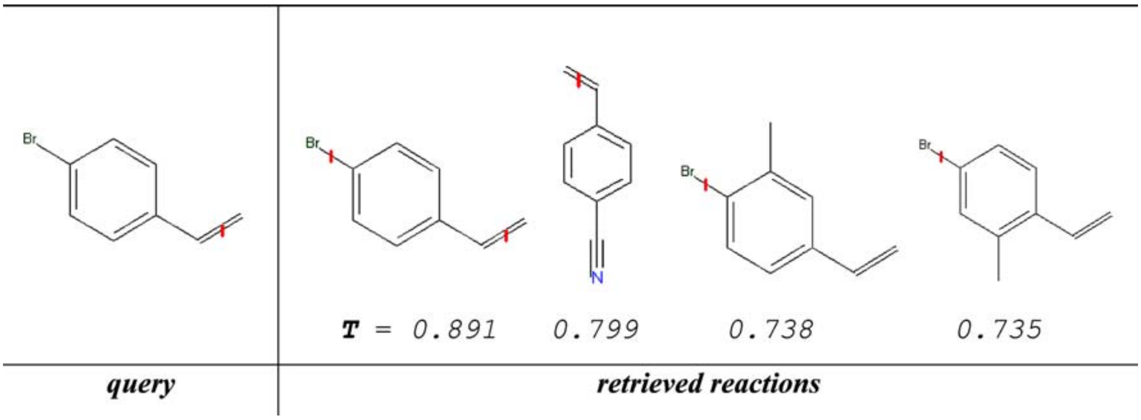


Figure 8. Similarity searching for a chemical reaction represented by a CGR: reactions retrieved in the Novalyst Discovery database using Tanimoto similarity threshold T_0 = 0.70. Tanimoto coefficients (T) were calculated using the descriptors pool involving both **I**(**AB**,2–6) sequences and **II**(Hy) augmented atoms. Dynamic bonds "╪" and "┼" correspond respectively to the transformation of a double bond to a single bond and to the breaking of a single bond (see Table 1).

Conclusions

It has been shown that substructural fragments represent a simple and safe way to encode conventional molecular structures as well as complex supramolecular systems and chemical reactions. The fragments can be easily calculated from 2D graphs, then be used as descriptors in structure – property studies or in similarity searching. The density of encoded information can be increased by reducing the number of selected fragment descriptors, which, however, leads to a reduction of the quality of related QSPR models.

Encoding chemical systems in descriptors/properties or descriptors/knowledge matrices provides the industrial and scientific communities with a rather secure way to share the data without revealing the structures, thus strengthening a partnership between them.

Acknowledgement

GDR PARIS and GDRE SupraChem are acknowledged for support. FH thanks Novalyst Discovery for a PhD fellowship.

References

1. Faulon, J.-L., Churchwell, C.J. and Visco, D.P. Jr., J. Chem. Inf. Comput. Sci., 43 (2003) 721.
2. Rucker, G. and Rucker, C., J. Chem. Inf. Comput. Sci., 41 (2001) 1457.
3. Churchwell, C.J., Rinoul, M.D., Martin, S., Visco, D.P. Jr., Kotu, A., Larson, R.S., Sillerud, L.O., Brown, D.C. and Faulon, J.-L., J. Mol. Graph. Mod., 22 (2004) 263.
4. Klopman, G. and Tu, M., J. Med. Chem., 42 (1999) 992.
5. Solov'ev, V.P., Varnek, A. and Wipff, G., J. Chem. Inf. Comput. Sci., 40 (2000) 847.
6. Klopman, G. and Zhu, H., J. Chem. Inf. Comput. Sci., 41 (2001) 439.
7. Varnek, A., Wipff, G. and Solov'ev, V.P., Solvent Extr. Ion Exch., 19 (2001) 791.
8. Artemenko, N.V., Baskin, I.I., Palyulin, V.A. and Zefirov, N.S., Doklady Akademii Nauk, 381 (2001) 317.
9. Varnek, A., Wipff, G., Solov'ev, V.P. and Solotnov, A.F., J. Chem. Inf. Comput. Sci., 42 (2002) 812.
10. Zefirov, N.S. and Palyulin, V.A., J. Chem. Inf. Comput. Sci., 42 (2002) 1112.
11. Solov'ev, V.P. and Varnek, A., J. Chem. Inf. Comput. Sci., 43 (2003) 1703.
12. Varnek, A., Fourches, D., Solov'ev, V.P., Baulin, V.E., Turanov, A. and Katritzky, A.R., J. Chem. Inf. Comput. Sci., 44 (2004) 1365.
13. Katritzky, A.R., Fara, D.C., Yang, H., Karelson, M., Suzuki, T., Solov'ev, V.P. and Varnek, A., J. Chem. Inf. Comput. Sci., 44 (2004) 529.
14. Clark, M., J. Chem. Inf. Model., 42 (2005) 30.
15. Faulon, J.-L., Collins, M.J. and Carr, R.D., J. Chem. Inf. Comput. Sci., 44 (2004) 427.
16. Faulon, J.-L., Visco, D.P. and Pophale, R.S., J. Chem. Inf. Comput. Sci., 43 (2003) 707.
17. Baskin, I., Skvortsova, M., Stankevich, I. and Zefirov, N., J. Chem. Inf. Comput. Sci., 35 (1995) 527.
18. Skvortsova, M., Baskin, I., Skvortsova, L., Palyulin, V., Stankevich, I. and Zefirov, N., Theochem: J. Mol. Struct., 466 (1999) 211.
19. Trepalin S.V., Gerasimenko V.A., Kozyukov A.V., Savchuk N. and Ph. Ivaschenko A.A., J. Chem. Inf. Comput. Sci., 42 (2002) 249.
20. Mavrovouniotis, M.L., Biotech. Bioeng., 36 (1990) 1070.
21. Mavrovouniotis, M.L., J. Biol. Chem., 266 (1991) 14440.
22. Meylan, W.M. and Howard, P.H., J. Pharm. Sci., 84 (1995) 83.

23. Hansch C. and Leo A., Exploring QSAR Fundamentals and Applications in Chemistry and Biology. ACS Prof. Ref. Book, Washington, 1995, 557 pp.
24. Klopman, G., Ding, C. and Macina, O.T., J. Chem. Inf. Comput. Sci., 37 (1997) 569.
25. Wang, R., Fu, Y. and Lai, L., J. Chem. Inf. Comput. Sci., 37 (1997) 615.
26. Golbraikh, A. and Tropsha, A., J. Chem. Inf. Comput. Sci., 43 (2003) 144.
27. Zheng, W. and Tropsha, A., J. Chem. Inf. Comput. Sci., 40 (2000) 185.
28. Dalby, A., Nourse, J.G., Hounshell, W.D., Gushurst, A.K.I., Grier, D.L., Leland, B.A. and Laufer, J., J. Chem. Inf. Comput. Sci., 32 (1992) 244.
29. Hanser T., Jauffret P., Marchaland J.F., Ellermann L., Gruber E., Kaufmann G., Am. Inst. Phys., Conference Proceeding no. 330, 1995, 575.
30. Jauffret P., Vogel H., Schildknecht S. and Kaufmann G., Learning synthetic knowledge from reaction databases: dealing with experimental conditions. Pub Informatics Ltd. Tetbury (Eng.), 2000.
31. Fujita, S., J. Chem. Inf. Comput. Sci., 26 (1986) 205.
32. Fujita, S., J. Chem. Inf. Comput. Sci., 27 (1987) 99.
33. http://infochimie.u-strasbg.fr/recherche/isida/index.php.
34. Korn G.A. and Korn T.M., Mathematical Handbook for Scientists and Engineers, 2nd Edition. McGraw-Hill Book Co, New York, 1968.
35. Golub, G.H. and Reinsch, C., Numer. Math., 14 (1970) 403.
36. Muller P.H., Neumann P. and Storm R., Tafeln der mathematischen Statistik VEB Fachbuchverlag: Leipzig, 1979, 280 pp.
37. Hou, T., Xia, K., Zhang, W. and Xu, X., J. Chem. Inf. Comput. Sci., 44 (2004) 266.
38. Bergström, C., Wassvik, C., Norinder, U., Luthman, K. and Artursson, P., J. Chem. Inf. Comp. Sci., 44 (2004) 1477.
39. Yan, A. and Gasteiger, J., J. Chem. Inf. Comput. Sci., 43 (2003) 429.
40. Delaney, J., J. Chem. Inf. Comput. Sci., 44 (2004) 1000.
41. Butina, D. and Gola, J., Chem. Inf. Comput. Sci., 43 (2003) 837.
42. Cheng, A. and Merz, K., J. Med. Chem., 46 (2003) 3572.
43. Engkvist, O. and Wrede, P., J. Chem. Inf. Comput. Sci., 42 (2002) 1247.
44. Klopman, G. and Zhu, H., J. Chem. Inf. Comput. Sci., 41 (2001) 439.
45. Lipinski, C., Lombardo, F., Dominy, B. and Feeney, P., Adv. Drug Delivery Rev., 46 (2001) 3.
46. Catana, C., Gao, H., Orrenius, C. and Stouten, P., J. Chem. Inf. Model., 45 (2005) 170.
47. Delaney, J.S., Drug Discovery Today, 10 (2005) 289.
48. Yaffe, D., Cohen, Y., Espinosa, G., Arenas, A. and Giralt, F., J. Chem. Inf. Comput. Sci., 41 (2001) 1177.
49. Ran, Y., Jain, N. and Yalkowsky, S., J. Chem. Inf. Comput. Sci., 41 (2001) 1208.
50. McElroy, N. and Jurs, P., J. Chem. Inf. Comput. Sci., 41 (2001) 1237.
51. Tetko, I., Tanchuk, V., Kasheva, T. and Villa, A., J. Chem. Inf. Comput. Sci., 41 (2001) 1488.
52. Raevsky O.A., Solov'ev V.P. and Grigor'ev V.Y., VINITI Deposited Doc. No. 1001–V88 (1988) 83 pp.
53. Drago, R.S., Ferris, D.C. and Wong, N., J. Am. Chem. Soc., 112 (1990) 8953.
54. Drago, R.S., Dadmun, A.P. and Vogel, G.C., Inorg. Chem., 32 (1993) 2473.
55. Abraham, M.H., Chem. Soc. Rev., 22 (1993) 73.
56. Abraham, M.H., Grellier, P.L., Prior, D.V., Taft, R.W., Morris, J.J., Taylor, P.J., Laurence, C., Berthelot, M., Doherty, R.M. et al., J. Am. Chem. Soc., 110 (1988) 8534.
57. Abraham, M.H. and Platts, J.A., J. Org. Chem., 66 (2001) 3484.
58. Raevskii, O.A., Grigor'ev, V.Y., Solov'ev, V.P. and Martynov, I.V., Doklady Akademii Nauk SSSR, 299 (1988) 1433.
59. Raevskii, O.A., Grigor'ev, V.Y., Solov'ev, V.P. and Martynov, I.V., Doklady Akademii Nauk SSSR, 298 (1988) 1166.
60. Raevskii, O.A., Grigor'ev, V.Y. and Solov'ev, V.P., Khimiko-Farmatsevticheskii Zhurnal, 23 (1989) 1294.
61. Raevsky, O.A., J. Phys. Org. Chem., 10 (1997) 405.
62. http://www.novalyst.com/.