Molecular Field Topology Analysis Method in QSAR Studies of Organic Compounds

Vladimir A. Palyulin, Eugene V. Radchenko, and Nikolai S. Zefirov*

Department of Chemistry, Moscow State University, Moscow 119899 Russia

Received June 19, 1998

A new method of QSAR analysis for organic compounds, molecular field topology analysis (MFTA), is considered that involves the topological superposition of the training set structures and the construction of a molecular supergraph (MSG). This enables the creation of the uniform descriptor vectors based on the local physicochemical parameters (atom and bond properties) of the molecules. The application of this technique is illustrated by a number of examples, and its features are discussed. The MFTA is especially suitable for solving the problems where the analysis of three-dimensional structure is either unnecessary or complicated.

1. INTRODUCTION

Modern studies of quantitative structure−activity relationships (QSARs) for organic compounds seek to reveal the structural features responsible for the interaction of molecules of an active compound (ligand) with a biological target. From the viewpoint of both the ligand/target fit and the possible design of new structures, it is clear that location of these features with respect to the molecule is as important as their character. A number of approaches taking into account the location of such features based on the topological as well as the spatial representation of structures was suggested in literature. The three-dimensional QSAR techniques such as comparative molecular field analysis (CoMFA)[1] potentially can provide the model of the target and account for the effect of various molecular parameters on the activity. However, their application is often complicated by a number of problems.

The existing topological methods for biological activity modeling do not provide sufficient generality and convenience since taking into account the position of important features is difficult. A number of approaches to QSAR analysis at the topological level[2] imply the consideration of a "superstructure" or "hyperstructure", i.e., such a graph that any structure of the training set can be represented as its subgraph. This superstructure can be viewed as a "topological lattice" that enables constructing a uniform description of the structures in a series suitable for subsequent statistical analysis. Historically, one of the first important steps in this direction was made by Free and Wilson.[3] Their technique consists in the superposition of substituents (being considered as a whole, without any regard to their structure) providing a simplified superstructure for a set of molecules that allows one to determine the contributions of different substituents in different positions. DARC/PELCO methodology[4,5] applies a related concept to the so-called hyperstructure: the molecular graphs are represented as a combination of the focus (common structural fragment usually crucial for the activity under study) and the limited concentric ordered environment (ELCO), the contributions of different sites of a hyperstructure (atoms, bonds, cycles, and other structural features in different positions of the ELCO) into an activity being determined by means of regression analysis. It is also possible to consider more complex descriptors accounting for the interactions and similarities in the effect of different sites. However, the extrapolation of the results to somewhat different structures might be complicated since the method fails to reveal the local physicochemical properties controlling the activity.

Another superstructure-based approach[6] which was applied to the recognition of activity types for a set of compounds consists of manual superposition of the structures in a chemically consistent way followed by the factor analysis of the atomic increments of molecular refraction for a number of positions.

The minimal topological difference (MTD) approach[7] to the active site mapping also involves the construction of a hypermolecule. Then, the initial map of the site is defined by assigning the values of -1, $+1$, and 0 to the hypermolecule vertices that are respectively (1) occupied in a standard compound, (2) unoccupied in a standard compound, and (3) indifferent for activity. The MTD is defined as a sum of these values for the vertices occupied in a given compound and of the number of -1 values in the map. Subsequent optimization of the values allows one to increase the correlation between the MTD and activity, thus providing the map of a target. Alternatively, the occupancy factors for the hypermolecule (after collapsing the identically occupied positions) may be used as descriptors for the direct regression analysis.[8]

The positional analysis technique[9] significantly develops these approaches by suggesting the use of the atomic physicochemical parameters that influence ligand binding to the target as descriptors (bond properties are not considered). This method was successfully applied to the analysis of various biological activities, demonstrating the viability of the topological approach taking into account the physicochemical data.[9,10] However, in all the examples these parameters were included only for some sites of a hypermolecule even when certain variation of atom types in other sites did exist. Moreover, only the substituent groups were considered, while the common fragment was ignored. When several possible superpositions existed for a given site, they

10.1021/ci980114i CCC: $19.00
Published on Web 03/14/2000

were selected by means of a formal protocol based on the relative size of the groups, even though some other structural features (atomic charges, H-bonding ability, etc.) might actually influence the binding of the molecule to a receptor. Also, the use of multiple linear regression method complicates the correct analysis of the relationships between the activity and descriptors for different sites.

On the other hand, the most widely used three-dimensional (3D) QSAR technique, comparative molecular field analysis (CoMFA),[1] is aimed at the detection of the spatial regions having positive or negative effect on activity with respect to various properties. Its procedure involves the spatial alignment of a series of structures followed by the calculation of steric, electrostatic, lipophilic, and other "potentials" (3D molecular fields) in the intersections of a three-dimensional lattice. After the statistical treatment of the resulting descriptor matrix by means of partial least squares (PLS) regression, one can obtain a structure−activity model and estimate the effect of various descriptors at different spatial regions on the activity. The use of descriptor selection procedures such as GOLPE[11,12] often increases the stability, predictivity, and interpretability of the model. However, the application of this method is complicated by a great number of descriptors, as well as by the alignment problem,[13] especially for conformationally flexible molecules where an induced fit between the ligand and the target may lead to ligand conformations substantially different from those optimal for the isolated molecule. Several techniques were devised in order to reduce the model sensitivity to alignment or altogether eliminate this stage of analysis. For instance, the steric field may be characterized by van der Waals envelope intersection volumes between a probe and the ligand molecule that provide smoother distance dependence than is the case for the standard Lennard-Jones 6−12 potential.[14] For a congeneric series of compounds, the canonical (formal) alignment rules are often necessary and sufficient to obtain the useful results.[13,15] The consideration of the spatial autocorrelation vectors[16] and the moments of mass and charge distributions[17] yields the molecular 3D descriptors that are invariant to both translation and rotation.

2. THE METHOD OF MOLECULAR FIELD TOPOLOGY ANALYSIS

Basic Principles. It is clear that in some cases the use of a topological rather than spatial alignment makes it possible to alleviate the problems typical of the CoMFA and other 3D-QSAR techniques. The quantitative description of structural features can be provided by local physicochemical parameters. One would expect a complementarity to exist between the distribution of these parameters for active compounds and the corresponding features in the target structure. In general, interaction cannot be attributed to a small number of key sites in a ligand molecule due to the possibility of an induced fit between the ligand and the receptor as well as a correlation between the parameters (especially electrostatic ones) of the neighboring sites.

To minimize these difficulties, we suggest, as a generalization of the approaches reviewed above, the method of molecular field topology analysis (MFTA), which may be considered as a "topological analog" of the CoMFA method. The overall process flow in this approach[18,19] is represented by Figure 1. First, for a set of structures with known activities (a training set) the so-called molecular supergraph (MSG) is automatically constructed. The MSG is a certain (not necessarily minimal or unique) graph such that each of the training set structures can be represented as its subgraph. For instance, Figure 2 shows one possible MSG **6** for a hypothetical series of structures **1**−**5**. The MSG enables the construction of the uniform descriptor vectors for all structures in the set. To construct each vector, the MSG vertices and edges corresponding to the atoms and bonds, respectively, of a given structure are assigned the values of local descriptors (e.g., atomic charge *Q* and van der Waals radius *R*) for these atoms and bonds, and unoccupied vertices and edges of the MSG are labeled with neutral descriptor values that provide a reasonable simulation of properties in respective unoccupied regions of space near a molecule. Thus, they should not be considered as "missing values" in a statistical sense. These parameters can also be optimized for some particular cases (similar to the dielectric constant in molecular mechanics computations). Nevertheless, our experience indicates that the structure−activity models are not very sensitive to their values within a reasonable range. The descriptor vector formation is illustrated in Figure 3. Generally, hydrogen atoms are not considered; however, the descriptor values on both the atom itself and the atoms attached to it (including hydrogens) are used in the analysis.

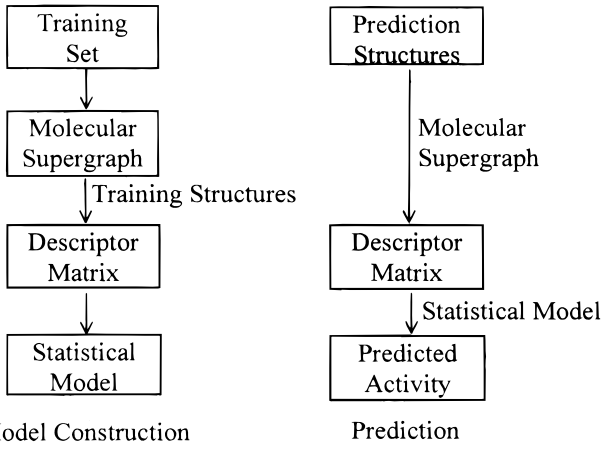


Figure 1. General procedure of model construction and activity prediction in the framework of MFTA approach.

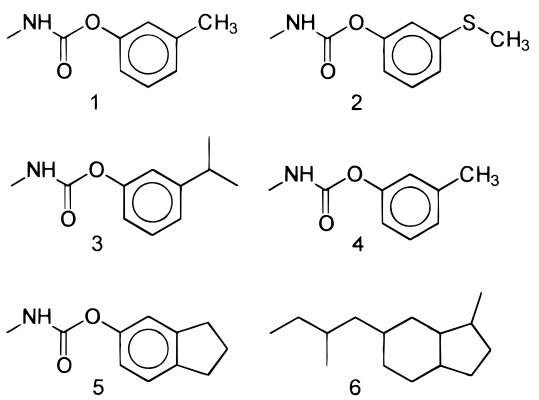


Figure 2. Construction of a molecular supergraph. Compounds **1**−**5** form a hypothetical training set. The minimal MSG for it is represented by graph **6**.

MSG Construction and Analysis. The mapping procedure used for generating the MSG and forming the descriptor

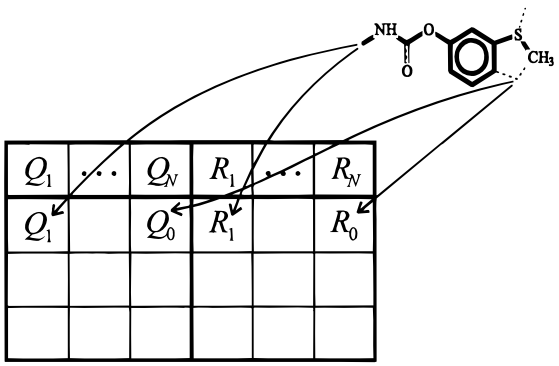


Figure 3. Descriptor vector formation. The structure of compound **2** (see Figure 2) is superimposed onto the MSG **6** (dashed line). The first MSG vertex maps to the structure atom (*N*-Me group), and the respective cells of the descriptor table contain the properties for that atom (atomic charge Q and van der Waals radius R). If MSG vertex has no corresponding atom in a given structure, the neutral descriptor values (Q_0 and R_0) are used.

vector is sufficiently versatile and allows for the consideration of the types of atoms, their valences, stereochemistry, bond types and orders, special limitations imposed by a researcher, and similarities in the distribution of local properties (electronegativity, atomic charge, H-bonding parameters, etc.) over the structure. Therefore, the resulting maps might be expected to reflect the anticipated character of the ligand−target interaction. The algorithm of mapping (in terms of graph theory, the maps correspond to the intersections of the MSG and a molecular graph) combines the algorithm of vertex-by-vertex expansion and the algorithm of searching for maximum cliques (complete subgraphs) of the module graph product[20] and efficiently finds the maximum connected graph intersections.

It may be formulated as follows. First of all, for two graphs U and V the graph of correspondence (compatibility) G^{UV} is constructed whose vertices $g_{ij}^{UV} = (u_i, v_j)$ specify the possible pair correspondences for the vertices of the initial graphs. Vertices $g_{ij}^{UV} = (u_i, v_j)$ and $g_{kl}^{UV} = (u_k, v_l)$ are adjacent only if edges (u_i, u_k) of graph U and (v_j, v_l) of graph V can correspond to each other. Therefore, the connected subgraph in G^{UV} that does not include the vertices with the coinciding u_i or v_j values gives rise to the connected intersection of graphs U and V. To enumerate such subgraphs, the recursive decomposition of the graph G^{UV} is used. At each decomposition step, among several possible variants of correspondence (graph G^{UV} vertices which can occur in the subgraph to be determined), preference is given to a vertex g_{ij}^{UV} that provides the maximum similarity in the distribution of local properties in the nearest environment of the vertex u_i of graph U and the vertex v_j of graph V. The definition of the dissimilarity measure is given in the Appendix. A nondeterministic (genetic) algorithm similar to that described in ref 21 was also implemented but proved to be inferior to the approach described above in terms of both computational speed and quality of intersections found.

During the construction of the MSG, the structures from the training set are processed sequentially. At each step, the intersection between the MSG constructed by this time (originally empty) and the next structure of the set is determined. Then, the MSG is augmented by the atoms that do not occur in the intersection, and the values of local properties are updated for all the MSG vertices. If necessary, all possible intersections may be considered. A similar search for intersections is performed during the formation of the descriptor vector. One can use the first (locally optimal) of several mappings, choose the mapping providing the maximum closeness of the resulting descriptors to the reference (most active) structure, or average the set of the descriptors with regard to the weights reflecting this closeness.

Since the number of descriptors is rather large (though much smaller than in CoMFA), the partial least squares (PLS) regression[22] is used to analyze the descriptor−activity relationships. As a result, the quantitative characteristic of the influence of each descriptor in each position, including common structural fragments, on activity can be determined. This enables the selection of the structural features controlling the activity with respect to their influence on the model predictivity or to the relative descriptor contribution into the model output defined by

$$I_j = \frac{b_j \,\mathrm{range}(x_j)}{\mathrm{range}(y)} \qquad (1)$$

where b_j is the coefficient at the descriptor x_j in the back-rotated PLS model, range(x_j) is the range of x_j, and range(y) is the range of activity in the training set.

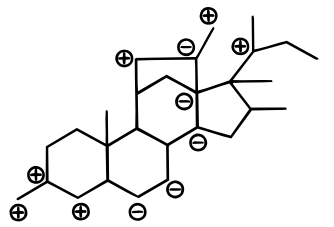
These structural elements can provide a basis for designing new, potentially more active structures as well as the anchor points for spatial structure alignment. Alternatively to PLS, the artificial neural networks may be applied to correlating the activity with descriptors.

Descriptors. To analyze the structure−activity relationships, the following descriptors are currently calculated: Gasteiger's atomic charge Q estimated with the electronegativity equalization approach,[23] Sanderson's electronegativity χ,[24] Bondi's van der Waals radius R,[25] atomic contribution to the molecular van der Waals surface S (the surface of the atom's van der Waals sphere excluding the areas intersected by other atoms; ternary intersections are neglected), relative steric accessibility defined as $A = S/S_{free}$ (where S_{free} is the van der Waals surface of the "free" (isolated) atom of the same type), electrotopological state ETS,[26] atomic lipophilicity contribution L_a taking into account the environment of an atom,[27] and group lipophilicity L_g defined as a sum of contributions for both a non-hydrogen atom and attached hydrogens, the ability of an atom in a given environment to be a donor (H_d) and acceptor (H_a) of a hydrogen bond characterized by the binding constants,[28] local stereochemical indicator variables, and the site occupancy factors for atoms P_a and bonds P_b (which have the value 1 if a given feature is present in the structure and 0 otherwise). This set of local descriptors provides sufficient coverage of major interaction types that are important for the interaction of ligand with a biological target. However, the set is open and can be easily extended to account for the specific features of the problem.

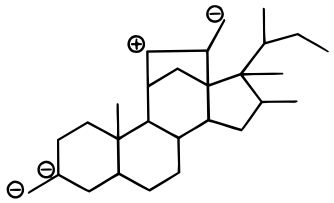
Implementation. The suggested method was implemented in a C++ computer program that calculates the local descriptors, builds the MSG, maps the descriptor values during model building and activity prediction, and constructs the MFTA structure−activity model.[29]

3. RESULTS AND DISCUSSION

The application of the method can be illustrated by the following examples.



Q

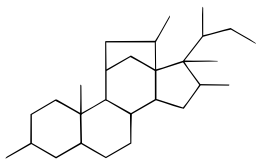


R

Figure 4. Modeling of the binding ability of steroids to corticosteroid-binding globulin. The major local descriptor contributions to activity (*Q*, atomic charge; *R*, van der Waals radius).

Steroid Binding to Corticosteroid-Binding Globulin. The activity of compounds[1] is characterized by the logarithm of the binding constant log K_b.[30] In the CoMFA method, the optimal model for the training set of 21 compounds contains two PLS factors and provides[1] the correlation coefficient $R = 0.947$ at the cross-validation parameter $Q^2 = 0.662$. However, the prediction quality for the test set of 10 compounds is rather low ($R = 0.225$, average error $\Delta_{av} = 0.611$). It can be somewhat increased by the exclusion of three compounds structurally dissimilar to those in the training set ($R = 0.341$, average error $\Delta_{av} = 0.506$).

In terms of the MFTA method, we have constructed the MSG **7**.



7

When using atomic charges and radii (neutral values 0) as descriptors, the optimal predictivity was obtained at the number of PLS factors $N_F = 2$ ($Q^2 = 0.715$, $R = 0.959$). Unless stated otherwise, the cross-validation computations were performed for four equal-sized groups of compounds (that is, at each step the data for 25% of compounds were excluded from model building, and then the error of prediction for these compounds was calculated). The major contributions of the local descriptors (relative contribution $|I_j| > 0.04$, see eq 1) to the model are shown in Figure 4. Figure 5 shows the predicted and experimental activity values for the test set obtained with MFTA ($R = 0.934$, $\Delta_{av} = 0.207$) and CoMFA. The results of prediction by both methods are listed in Table 1. It should be noted that the results of CoMFA and MFTA analyses are not directly comparable due to the different nature of descriptors as well as the limitations inherent in the back-rotation of the PLS model. For instance, some of the most important areas of steric interaction (substituents in positions 3 and 11 of the steroid skeleton) coincide in these methods while the effect of some other areas is different. Also, contrary to the findings of

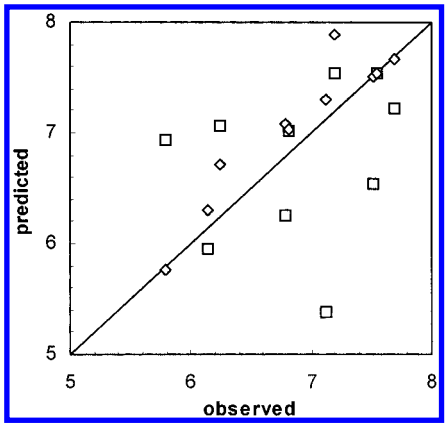
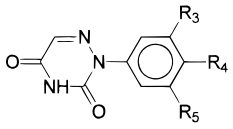


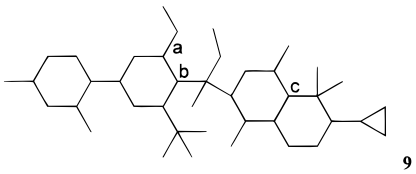
Figure 5. Modeling of the binding ability of steroids to corticosteroid-binding globulin. Activity prediction for the test set by the CoMFA (□) and MFTA (◇) methods.

CoMFA method, the MFTA reveals a considerable contribution of electrostatic parameters to the model (the exclusion of the charge descriptors leads to the moderate decrease in modeling quality and the dramatic decrease in predictivity).

Anticoccidial Activity of Substituted Triazinediones. The compounds of type **8** were characterized by the logarithm of the minimal efficient concentration log(1/MEC).[15]



8



9

The optimal model for the training set of 54 compounds ($N_F = 4$) includes charges, radii, and van der Waals surfaces of atoms (neutral values 0) as descriptors. The most active compound, diclazuril ($R_3 = R_5 = Cl$, $R_4 = CH(CN)$-*p*-C_6H_4-Cl), was used as a reference. Figure 6 shows the MSG structure and the major contributions of local descriptors ($|I_j| > 0.03$) to the activity. The characteristics of the models constructed by different methods (Table 2) and the comparison of the predicted activities with experimental values for the test set of 14 compounds (Figure 7) demonstrates that the MFTA method gives a more adequate model. One can note that the relative importance of various parameters differs from that of CoMFA model in ref 15. However, both approaches identify the similar steric effects (contributions of molecular surface descriptor and steric potential in MFTA and CoMFA, respectively) in position R_3 of the common phenyl group (marked with *a* in the MSG **9**) and in position marked with *c*. Also, the two models agree in indicating the preference for a positive charge near position R_4 of the common phenyl group (marked with *b*) and for a negative charge near position *c* (MSG **9**). In addition, the MFTA model reveals some other correlations between the local properties and activity.

Inhibition of Acetylcholinesterase by Monosubstituted Phenyl-*N*-methylcarbamates. One hundred twenty-four

Table 1. Prediction Results for the Corticosteroid-Binding Ability of Steroids by Means of the MFTA Method

name	log K_b obsd	log K_b pred
11β,17,21-trihydroxy-1,4-pregnadiene-3,20-dione (prednisolone)	7.512	7.505
21-acetoxy-11β,17-dihydroxy-4-pregnene-3,20-dione (cortisol 21-acetate)	7.553	7.538
4-pregnene-3,11,20-trione	6.779	7.081
11α,21-dihydroxy-4-pregnene-3,20-dione (epicorticosterone)	7.200	7.884
17β-hydroxy-4-estren-3-one (19-nortestosterone)	6.144	6.295
16α,17-dihydroxy-4-pregnene-3,20-dione	6.247	6.712
16α-methyl-4-pregnene-3,20-dione	7.120	7.306
19-nor-4-pregnene-3,20-dione (19-norprogesterone)	6.817	7.027
11β,17,21-trihydroxy-2α-methyl-4-pregnene-3,20-dione	7.688	7.664
11β,17,21-trihydroxy-2α-methyl-9α-fluoro-4-pregnene-3,20-dione	5.797	5.767

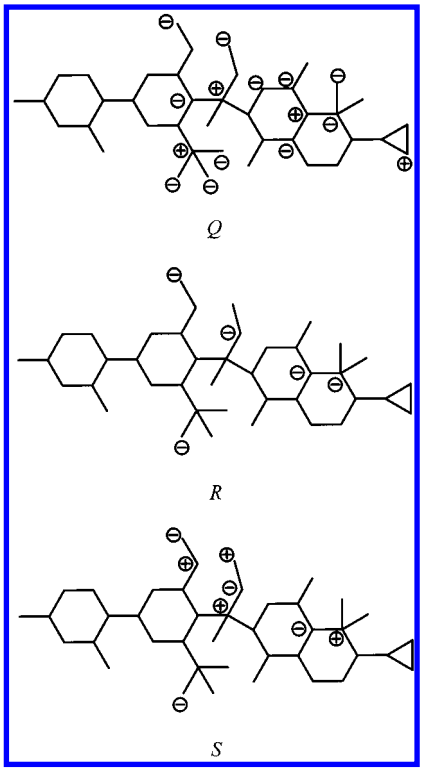
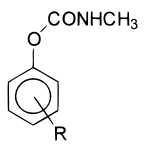


Figure 6. Modeling of the anticoccidial activity of substituted triazinediones. The major local descriptor contributions to activity (*Q*, atomic charge; *R*, van der Waals radius; *S*, van der Waals surface contribution).

Table 2. Statistical Characteristics of Different Models of Anticoccidial Activity of Triazinediones (*R*, Correlation Coefficient; Q^2, Cross-Validation Parameter; Δ_{av}, Average Error)

	training set		test set	
descriptors	*R*	Q^2	*R*	Δ_{av}
physicochemical[15]	0.749	0.488	0.740	0.73
CoMFA[15]	0.811	0.469	0.362	0.95
CoMFA and physicochemical[15]	0.895	0.613	0.671	0.71
structural fragments[34]	0.900		0.621	0.80
MFTA	0.922	0.666	0.812	0.56

compounds of type **10** were characterized by their pI_{50} values.[9] The optimal model ($N_F = 6$) includes charges (neutral value 0), lipophilicity (neutral value −0.25), and



10

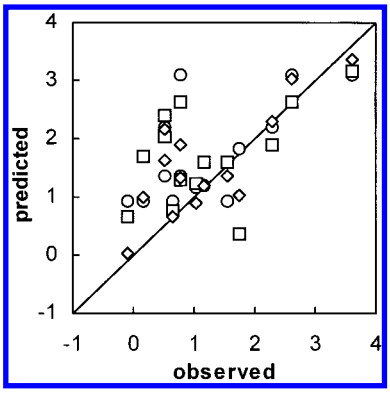


Figure 7. Modeling of the anticoccidial activity of substituted triazinediones. Activity prediction for the test set by means of the fragment scheme (○), CoMFA (□), and MFTA (◇).

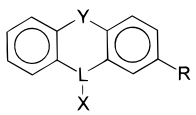
Table 3. Statistical Characteristics of Different Models for the Training Set of Phenyl-*N*-methylcarbamates, Inhibitors of Acetylcholinesterase, Based on the Positional Analysis[9] and MFTA Methods (*n*, Number of Compounds; *s*, Standard Deviation; Other Notations, Table 2)

method	*n*	*R*	Q^2	*s*	Δ_{av}
pos. analysis	46 (ortho)	0.829	0.493	0.486	0.335
pos. analysis	36 (meta)	0.841	0.674	0.390	0.297
pos. analysis	40 (para)	0.865	0.633	0.341	0.265
MFTA	124	0.963	0.666	0.270	0.192

hydrogen-bond acceptor ability (neutral value −2) of atoms as descriptors. Table 3 lists the model parameters for the training set.

The MSG structure and the major descriptor contributions ($|I_j| > 0.05$) to activity are shown in Figure 8. Contrary to ref 9, we described all the compounds of the series in terms of one model and revealed a certain similarity in the influence of substituents in different positions.

Tumor Multidrug Resistance (MDR) Reversing Activity of Phenothiazines and Thioxanthenes. Nineteen compounds of type **11** were characterized[31] by their log (MDR



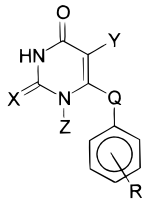
11

ratio) values (MDR ratio represents the decrease of IC_{50} in doxorubicin treatment; only the more potent trans isomers were considered).

The optimal predictivity ($N_F = 3$, $R = 0.989$, $Q^2 = 0.676$) is obtained for the model including atomic charges, radii, and bond presence parameters (neutral values 0). The MSG and the most important descriptors ($|I_j| > 0.035$) are shown

in Figure 9. These results confirm some correlations made in ref 31 with the type of structures but also reveal certain other factors which influence the activity. In particular, it should be noted that model quality decreases if one does not consider the bond descriptor P_b.

Anti-HIV-1 Activity of Substituted Benzylpyrimidines. The set of 84 compounds[32] of type **12** was characterized by the $\log(1/EC_{50})$ values.



12

The optimal model ($N_F = 5$, $R = 0.954$, $Q^2 = 0.737$, six cross-validation groups) includes radius (neutral value 0), lipophilicity (neutral value −0.25), and hydrogen-bond acceptor ability of atoms (neutral value −2) as descriptors. The MSG structure and the most important descriptor contributions ($|I_j| > 0.025$) are shown in Figure 10. Our results support the earlier conclusion[32] that the activity is enhanced by the hydrophobic substituents. However, some other factors are also revealed by the MFTA analysis.

For this activity, the design of new, potentially more active structures was attempted. Taking into account the structural variation within the training set and the effect of descriptors in different positions, we have determined a set of substituents and performed the exhaustive structure generation[33] for this series of compounds using the following substituents: X/Q = O/CH_2, O/S, S/S; Y = H, Me, Pr, ^{i}Pr, ^{c}Pr, allyl; Z = H, Bu, CH_2OEt, $CH_2OCH_2CH_2OH$, CH_2OCH_2-Ph, $CH_2OCH_2{}^c$Hex; R = H, 3-Me, 3-OMe, 3-Cl, 3-CN, 3-Ac, 5-Me, 5-Cl. Then the activity values for each of the 1125 structures were predicted using the MFTA model, and the 10 best structures are listed in Table 4. The comparison of predicted and observed values shows that we were able to reproduce the most potent compound of the original data set as well as to suggest several new promising structures.

4. CONCLUSION

We have described a method of molecular field topology analysis (MFTA) based on constructing the molecular supergraph and analyzing the local physicochemical characteristics of structures (atomic and bond parameters), which allows one to reveal the relationships between the biological activity of organic compounds and their structural features. MFTA often gives the models that are comparable or superior in quality of description and prediction to the models based on the widely used classical QSAR methods and 3D approaches. Thus, this method should be regarded as complementing the existing techniques such as CoMFA. In particular, it is useful when the consideration of the 3D structure is either unnecessary or complicated.

ACKNOWLEDGMENT

We wish to express our gratitude to the Russian Foundation for Basic Research for the support of this work.

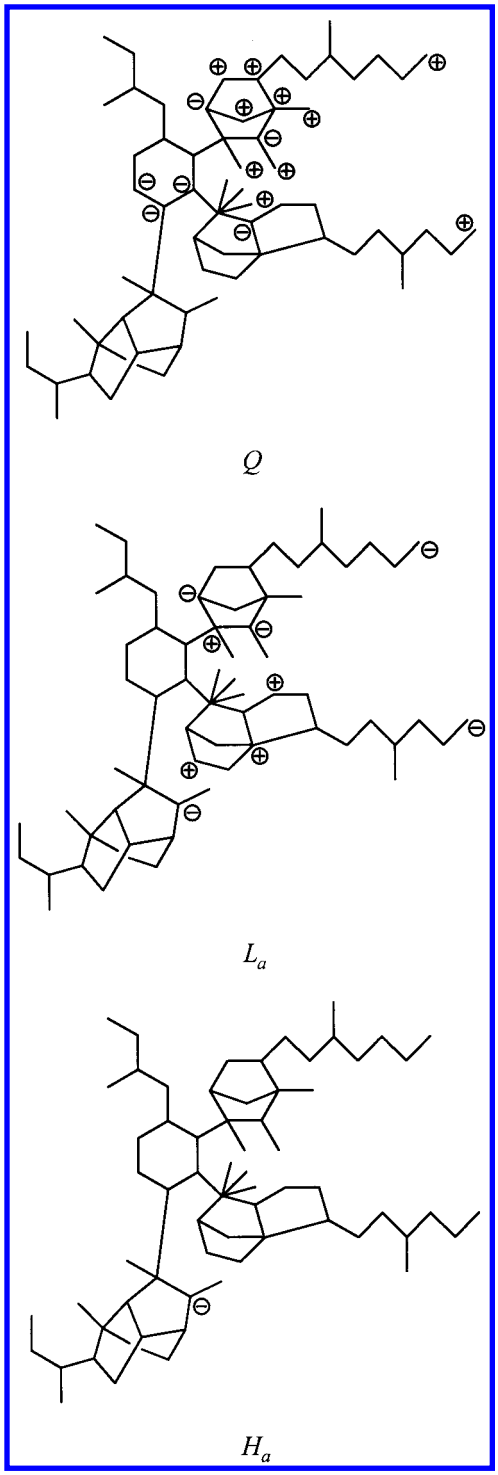


Figure 8. Major local descriptor contributions to acetylcholinesterase inhibition activity of phenyl-*N*-methylcarbamates (*Q*, atomic charge; L_a, atom lipophilicity contribution; H_a, hydrogen bond acceptor ability).

APPENDIX. ENVIRONMENT DISSIMILARITY MEASURE FOR GRAPH MAPPING

Consider two graphs whose vertices are characterized by a number of descriptors. We aim to construct a measure of descriptor value similarity for the environments of vertex u_i in graph U and vertex v_j in graph V. It should be noted that such traditional similarity measures of numerical data sets as correlation coefficient are not suitable for graph mapping, since the environments in different graphs may contain

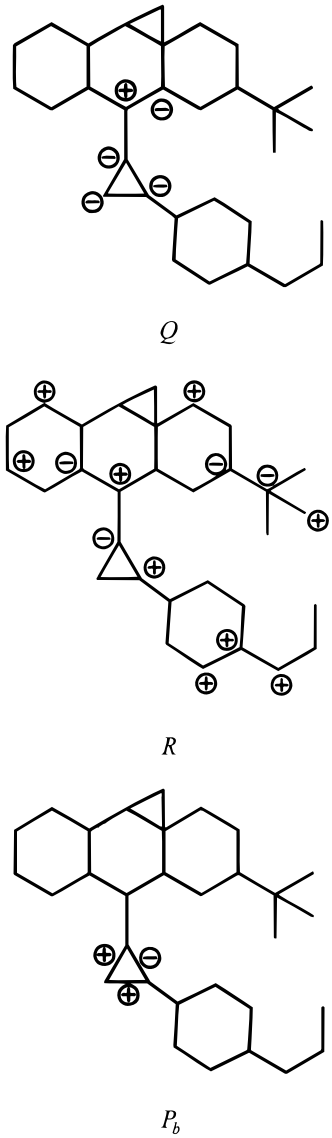


Figure 9. Major local descriptor contributions to MDR activity of phenothiazines and thioxanthenes (*Q*, atomic charge; *R*, van der Waals radius; P_b, bond occupancy factor).

different number of vertices, giving rise to the sets of different size. In this paper, the environment dissimilarity measure for the vertices u_i and v_j is defined as

$$D(i,j) = \sum_{l=0}^{K} \frac{\mathrm{LDiff}(l,i,j)}{F^l} \qquad (2)$$

where l is the current environment level, K is the highest level considered, F is the damping factor making the impact of differences decrease with distance, and the "distance" between the environments for level l is defined as

$$\mathrm{LDiff}(l,i,j) = \frac{1}{n(E_{li}^U) + n(E_{lj}^V)} \sum_d \frac{\mathrm{SDiff}(l,d,i,j)}{\sqrt{\mathrm{var}(X^{Ud})\,\mathrm{var}(X^{Vd})}}$$

X^{Ud} is the descriptor value set for descriptor d and graph U, $\mathrm{var}(X^{Ud})$ is the variance of X^{Ud}, E_{li}^U is the l-order environment of vertex i in graph U, $n(E_{li}^U)$ is the number of vertices in E_{li}^U, X^{EUd} is the descriptor value set for E_{li}^U and descriptor

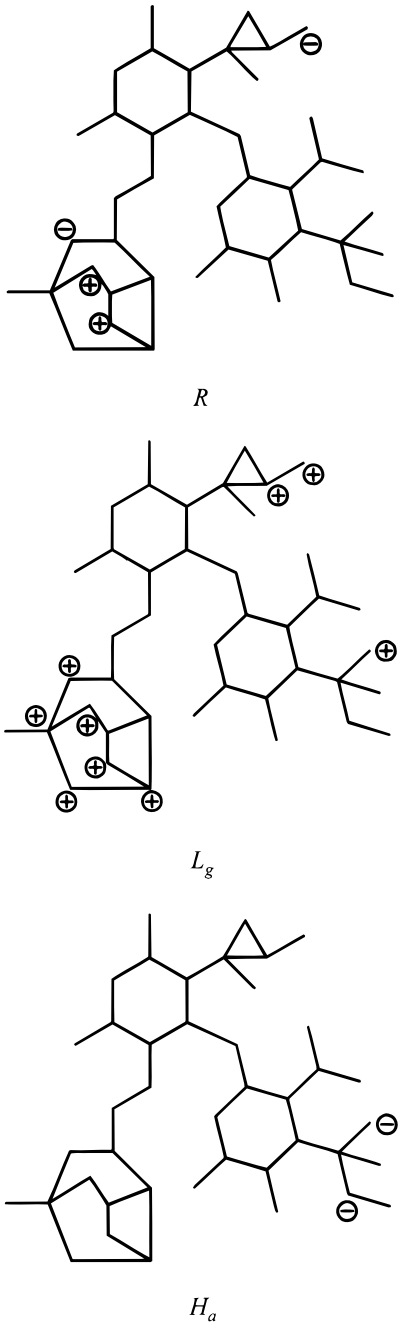


Figure 10. Major local descriptor contributions to anti-HIV-1 activity of benzylpyrimidines (*R*, van der Waals radius; L_g, group lipophilicity contribution; H_a, hydrogen bond acceptor ability).

Table 4. Design of New Potentially More Active Benzylpyrimidine Structures of Type **11** Based on the MFTA Model

no.	X	Q	Y	Z	R	pred	obsd
11−1	O	S	^{i}Pr	CH_2OCH_2Ph	3,5-Me_2	9.75	
11−2	O	CH_2	^{i}Pr	CH_2OCH_2Ph	3,5-Me_2	9.56	
11−3	S	S	^{i}Pr	CH_2OCH_2Ph	3,5-Me_2	9.50	
11−4	O	S	^{i}Pr	CH_2OEt	3,5-Me_2	9.43	
11−5	O	S	^{i}Pr	CH_2OCH_2Ph	3,5-Cl_2	9.36	
11−6	O	S	^{i}Pr	CH_2OCH_2Ph	3-Ac, 5-Cl	9.33	
11−7	O	CH_2	^{i}Pr	CH_2OEt	3,5-Me_2	9.25	9.22 (best)
11−8	O	S	^{i}Pr	CH_2OCH_2Ph	3-Me	9.25	
11−9	S	S	^{i}Pr	CH_2OEt	3,5-Me_2	9.18	
11−10	O	CH_2	^{i}Pr	CH_2OCH_2Ph	3,5-Cl_2	9.17	

d, and the "distance" $\mathrm{SDiff}(l,d,i,j)$ between the sets X^{EUd} and X^{EVd} is defined as

Table 5

	u_1	u_2	v_1	v_1	v_3
χ	2.746	3.475	2.746	2.746	4.000
R	1.70	1.75	1.70	1.70	1.47

Table 6

	u_1, v_1	u_1, v_2	u_1, v_3	u_2 ,v_1	u_2, v_2	u_2, v_3
D_χ	1.236	0.840	8.550	2.472	3.691	1.282
D_R	0.460	5.110	19.908	0.919	4.161	28.824

Table 7

	$\{v_1,v_2\}$	$\{v_2,v_1\}$	$\{v_2,v_3\}$	$\{v_3,v_2\}$
ΣD_χ	4.927	3.312	2.122	12.241
ΣD_R	4.621	6.029	33.934	24.069

$$\mathrm{SDiff}(l,d,i,j) = \sum_{p=1}^{n(E^U)} \min_q (X_p^{EUd} - X_q^{EVd})^2 + \sum_{q=1}^{n(E^V)} \min_p (X_p^{EUd} - X_q^{EVd})^2$$

If $\mathrm{var}(X^{Ud}) = 0$ for a given graph, the value 1 is used instead. This does not affect the applicability of results since no choice between U vertices exists in this case. If $n(E_{li}^{U})$ or $n(E_{lj}^{V})$ is equal to 0 for a given level l, the value LDiff($l - 1$,i,j) is used instead of LDiff(l,i,j) for the current and all subsequent levels.

The measure constructed in such a way provides intuitively desirable results even if the numbers of environment vertices in two graphs are different. For instance, consider the dissimilarity values ($K = 1$, $F = 2$) for the atoms of the following compounds with respect to electronegativity χ[24] and van der Waals radius R[25]:

U *V*

The descriptor values are listed in Table 5.

From these data, the dissimilarity values in Table 6 can be calculated for vertex pairs (u_i, v_j) with respect to both descriptors.

By summing up the vertex dissimilarity values, we can obtain the total dissimilarity associated with any given mapping of graph vertices $\{u_1, u_2\}$ (see Table 7).

Evidently, an optimal mapping with respect to electronegativity values is $\{u_1, u_2\} \leftrightarrow \{v_2, v_3\}$while that to the radii is $\{u_1, u_2\} \leftrightarrow \{v_1, v_2\}$. This example demonstrates that one cannot formulate a universal mapping criterion applicable in any situation. Instead, the mapping procedure should take into account the parameters important for the activity under study.

REFERENCES AND NOTES

(1) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(2) Kubinyi, H. *QSAR: Hansch Analysis and Related Approaches*; VCH: Weinheim, 1993.
(3) Free, S. M.; Wilson, J. M. A Mathematical Contribution to Structure−Activity Studies. *J. Med. Chem.* **1964**, *7*, 395−399.
(4) Dubois, J.-É.; Laurent, D.; Aranda, A. Système DARC. XVI. Théorie de Topologie-Information. I. Méthode de Perturbation d'Environnements Limités Concentriques Ordonnés (PELCO). *J. Chim. Phys.* **1973**, *70*, 1608−1615.
(5) Mercier, C.; Fabart, V.; Sobel, Y.; Dubois, J.-É. Modeling Alcohol Metabolism with the DARC/CALPHI System. *J. Med. Chem.* **1991**, *34*, 934−942.
(6) Menon, G. K.; Cammarata, A. Pattern Recognition II: Invesigation of Structure−Activity Relationships. *J. Pharm. Sci.* **1977**, *66*, 304−314.
(7) Simon, Z.; Badilescu, I.; Racovitan, T. Mapping of Dihydrofolate-reductase Receptor Site by Correlation with Minimal Topological (Steric) Differences. *J. Theor. Biol.* **1977**, *66*, 485−495.
(8) Simon, Z.; Holban, S.; Motoc, I. Steric Mapping of a Pancreatic Carboxypeptidase Inhibition Site and a Free-Wilson Procedure Based upon the Hypermolecule. *Rev. Roum. Biochim.* **1979**, *16*, 141−145.
(9) Magee, P. S. A New Approach to Active-Site Binding Analysis. Inhibitors of Acetylcholinesterase. *Quant. Struct.-Act. Relat.* **1990**, *9*, 202−215.
(10) Magee, P. S. Positional Analysis of Binding Events. In *QSAR: Rational Approaches to the Design of Bioactive Compounds*; Silipo, C., Vittoria, A., Eds.; Elsevier: Amsterdam, 1991.
(11) Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9−20.
(12) Pastor, M.; Cruciani, G.; Clementi, S. Smart Region Definition: A New Way to Improve the Predictive Ability and Interpretability of Three-Dimensional Quantitative Structure-Activity Relationships. *J. Med. Chem.* **1997**, *40*, 1455−1464.
(13) Cramer, R. D.; Clark, R. D.; Patterson, D. E.; Ferguson, A. M. Bioisosterism as a Molecular Diversity Descriptor: Steric Fields of Single "Topomeric" Conformers. *J. Med. Chem.* **1996**, *39*, 3060−3069.
(14) Sulea, T.; Oprea, T. I.; Muresan, S.; Chan, S. L. A Different Method for Steric Field Evaluation in CoMFA Improves Model Robustness. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1162−1170.
(15) McFarland, J. W. Comparative Molecular Field Analysis (CoMFA) of Anticoccidial Triazines. *J. Med. Chem.* **1992**, *35*, 2543−2550.
(16) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769−7775.
(17) Silverman, B. D.; Platt, D. E. Comparative Molecular Moment Analysis (CoMMA): 3D-QSAR without Molecular Superposition. *J. Med. Chem.* **1996**, *39*, 2129−2140.
(18) Radchenko, E. V.; Palyulin, V. A.; Zefirov, N. S. Molecular Field Topology Analysis: Local Physicochemical Descriptors in QSAR Studies of Organic Compounds. In *Eleventh European Symposium on Quantitative Structure-Activity Relationships: Computer-Assisted Lead Finding and Optimization*, Lausanne, Switzerland, Sept 1−6, 1996; P21A.
(19) Zefirov, N. S.; Palyulin, V. A.; Radchenko, E. V. Molecular Field Topology Analysis in Studies of Quantitative Structure-Activity Relationships for Organic Compounds. *Dokl. Chem.* **1997**, *352*, 23−26.
(20) Bessonov, Yu. E. On Solution of the Detection Problem for the Maximum Graph Intersections by Analyzing the Projections of Module Product Subgraphs. *Vychisl. Sist.* **1985**, *112*, 3−22 (Russian).
(21) Brown, R. D.; Jones, G.; Willett, P.; Glen, R. C. Matching Two-Dimensional Chemical Graphs Using Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 63−70.
(22) Martens, H.; Naes, T. *Multivariate Calibration*; Wiley: Chichester, 1989.
(23) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity: a Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3228.
(24) Sanderson, R. T. Electronegativity and Bond Energy. *J. Am. Chem. Soc.* **1983**, *105*, 2259−2261.
(25) Bondi, A. van der Waals Volumes and Radii. *J. Phys. Chem.* **1964**, *68*, 441−451.
(26) Hall, L. H.; Mohney, B.; Kier L. B. The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76−82.
(27) Ghose, A. K.; Pritchett, A.; Crippen, G. M. Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure−Activity Relationships III: Modeling Hydrophobic Interactions. *J. Comput. Chem.* **1988**, *9*, 80−90.
(28) Abraham, M. H.; Duce, P. P.; Prior, D. V.; Barratt, D. G.; Morris, J. J.; Taylor, P. J. Hydrogen Bonding. Part 9. Solute Proton-Donor and Proton-Acceptor Scales for Use in Drug Design. *J. Chem. Soc., Perkin Trans. 2* **1989**, *10*, 1355−1375.

(29) Additional information on the program and its applications is available from the authors (vap@org.chem.msu.su).
(30) The test and training set structures were corrected in accordance with the original publications: Dunn, J. F.; Nisula, B. C.; Rodbard, D. Transport of Steroid Hormones: Binding of 21 Endogenous Steroids to both Testosterone-Binding Globulin and Corticosteroid-Binding Globulin in Human Plasma. *J. Clin. Endocrin. Metab.* **1981**, *53*, 58−68. Mickelson, K. E.; Forsthoefel, J.; Westphal, U. Steroid-Protein Interactions. Human Corticosteroid Binding Globulin: Some Physicochemical Properties and Binding Specificity. *Biochemistry* **1981**, *20*, 6211−6218.
(31) Pajeva, I. K.; Wiese, M. QSAR and Molecular Modelling of Catamphiphilic Drugs Able to Modulate Multidrug Resistance in Tumors. *Quant. Struct.-Act. Relat.* **1997**, *16*, 1−10.
(32) Garg, R.; Kurup, A.; Gupta, S. P. Quantitative Structure-Activity Relationship Studies on Some Acyclouridine Derivatives Acting as Anti-HIV-1 Drugs. *Quant. Struct.-Act. Relat.* **1997**, *16*, 20−24.
(33) Tratch, S. S.; Lomova, O. A.; Sukhachev, D. V.; Palyulin, V. A.; Zefirov, N. S. Generation of Molecular Graphs for QSAR Studies: an Approach Based on Acyclic Fragment Combinations. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 130−139.
(34) Zefirov, N. S.; Petelin, D. E.; Palyulin, V. A.; McFarland, J. W. Quantitative Relationship between the Structure of 2-Substituted 1,2,4-Triazine-3,5(2H,4H)-diones and their Anticoccidial Activity. *Dokl. Akad. Nauk* **1992**, *327*, 504−508 (Russian).

CI980114I