The Comparative Molecular Surface Analysis (COMSA) − A Nongrid 3D QSAR Method by a Coupled Neural Network and PLS System: Predicting pK_a Values of Benzoic and Alkanoic Acids

Jarosław Polański,* Rafał Gieleciak, and Andrzej Bąk

Department of Organic Chemistry, Institute of Chemistry, University of Silesia, PL-40-006 Katowice, Poland

Received March 15, 2001

A self-organizing neural network was used to design a novel method capable of the quantitative prediction of molecular properties. The method is based on the comparison of molecular surfaces performed by the coupled neural network and PLS system. Unlike CoMFA and related methods it does not compare the properties describing a discrete set of points but the average property values calculated for a certain area of the molecular surface. It has been found that the results of the PLS analysis of the series of the comparative matrices of the molecular electrostatic potential (MEP) are quite stable. Also the results only slightly depend on such parameters as the number of points sampled at the molecular surface (D) or a winning distance (MD) of the self-organizing neurons. The influence of these parameters for modeling the effects limited by steric and electronic effects was determined and the pK_a values of the ortho-, meta-, and para- (o-, m-, p-) analogues of benzoic acid and selected alkanoic acids were predicted. We generally found that for the series analyzed CoMSA gave better models than CoMFA.

INTRODUCTION

Predicting molecular properties and modeling chemical or biological effects are one of the most challenging aims of present day chemistry and pharmacology. The cost of the technology involved in the development of a new drug means the trial and error strategy is unacceptable. The expansion of rational techniques, in particular the Quantitative Structure Activity Relationships (QSAR) and all its variants, i.e., QSPR, QSRR, and finally three-dimensional (3D) approaches[1−4] is obviously necessary.

3D QSAR strategies, especially CoMFA have notably contributed to our ability to forecast the activity of potential bioefectors. Theoretically, 3D methods should offer a clear advantage over 2D techniques. Practically, the classical Hansch approach still remains an important method.[5−7] A number of publications have shown that CoMFA protocols can be modified to provide models of the better predictive quality. This includes the improvements in the description of molecules, rules for the alignment of molecules as well as the statistics used. [8−11] As proper alignment of the molecules is an important problem in CoMFA various attempts have been made to improve that process. Among many suggested modifications in the alignment of molecules flexibility is a feature that can provide better description of the binding to the putative receptor.[11] One of the alternative potential uses in this field are neural networks[12,13] and in particular self-organizing neural networks.[14−16]

In the current work we discuss the results of our study aimed at the implementation of a coupled neural network and PLS system for designing a nongrid technique capable of analyzing the similarity of three-dimensional molecular surfaces, modeling 3D QSAR's, and predicting molecular properties.

Theoretical Background. Self-organizing neural network (SOM) is a technique designed to reduce the dimensionality of the data while preserving topology. In recent years this technique has been applied to the analysis of the chemical information.[15,16] The Kohonen SOM was used for the transformation of three-dimensional molecular surfaces into the two-dimensional maps of the molecular electrostatic potential.[16,17] The partial atomic charges of the atomic molecular representations were also projected into two-dimensional topographic maps.[18] It has been shown that there are some important properties of the Kohonen transformation. First, the ability to compress the size of the data and second to reconstruct the 3D object from the 2D representation. It is these abilities that makes this procedure an interesting tool for molecular design.[15,16,19] Such maps were used for the visualization of the interactions of individual molecules with biological receptors or designing drugs.[15,16]

EXPERIMENTAL SECTION

Model Building. All the experimental data, i.e., pK_a for the benzoic acids and the pK_a for the alkanoic acids, are extracted from the ref 20 and are given in Tables 1 and 2.

We used Gesteiger's software package for modeling purposes. The 3D-coordinates of all molecules were obtained by the 3D structure generator CORINA[21] from the constitution of the respective molecules.[22,23] Partial atomic charges were calculated by the PEOE method,[24,25] and the SURFACE program was used for the calculation of the Coulomb electrostatic potential on the molecular surface. We used both the neutral and anionic form of the acid molecules.

CoMFA Analysis. All modeling work in this part was carried out with the Sybyl 6.6 software package. Structure optimizations were performed with the Tripos standard molecular mechanics force field using the POWELL mini-

* Corresponding author e-mail: Polanski@us.edu.pl.

10.1021/ci010031t CCC: $22.00
Published on Web 02/26/2002

Table 1. Comparison of the Measured and Predicted pK_a Values for Benzoic Acids **1**−**41**

no.	benzene ring − substituent	p$K_{(exp)}$	p$K_{(pred)}$[a]	p$K_{(pred)}$ (SD)[b]	p$K_{(pred)}$[c]	p$K_{(pred)}$[d]
1	3 $-NH_2$	4.78	4.51	4.47 (0.06)	4.63	4.06
2	2 $-NH_2$	4.95	3.76	3.67 (0.10)	3.86	3.13
3	4 $-NH_2$	4.85	4.33	4.34 (0.08)	4.49	4.38
4	2 $-C(CH_3)_3$	3.54	3.53	3.34 (0.15)	3.75	3.80
5	3 −Br	3.90	3.97	4.00 (0.03)	4.13	3.91
6	2 −Br	2.85	3.20	3.35 (0.11)	2.95	3.27
7	4 −Br	3.97	4.13	4.18 (0.04)	4.20	4.05
8	4 $-C(CH_3)_3$	4.40	4.38	4.43 (0.04)	4.42	3.96
9	3 −Cl	3.82	3.79	3.86 (0.04)	3.96	3.76
10	2 −Cl	2.94	3.03	3.25 (0.14)	3.30	3.02
11	4 −Cl	3.99	3.99	4.03 (0.05)	4.03	4.11
12	3 −CN	3.60	3.35	3.28 (0.09)	3.56	3.56
13	4 −CN	3.55	3.53	3.53 (0.07)	3.60	3.55
14	2 −OH, 3 −OH	2.94	2.92	3.02 (0.10)	3.14	2.75
15	2 −OH, 4 −OH	3.29	3.35	3.27 (0.07)	2.92	4.04
16	2 −OH, 5 −OH	2.97	2.95	2.86 (0.11)	3.11	3.13
17	2 −OH, 6 −OH	1.30	1.35	1.49 (0.13)	2.49	1.96
18	3 −OH, 4 −OH	4.48	4.51	4.40 (0.09)	3.89	3.48
19	3 −OH, 5 −OH	4.04	4.11	4.21 (0.16)	4.07	4.41
20	2 $-C_2H_5$	3.79	3.92	3.91 (0.12)	3.80	4.32
21	4 $-C_2H_5$	4.35	4.36	4.44 (0.05)	4.42	4.58
22	3 −F	3.87	4.03	3.94 (0.09)	3.89	4.02
23	2 −F	3.27	3.32	3.49 (0.19)	3.51	3.48
24	4 −F	4.14	3.94	3.98 (0.06)	3.88	3.98
25	2 −OH	3.00	2.95	2.89 (0.11)	3.24	3.04
26	4 −OH	4.58	4.65	4.50 (0.12)	4.14	4.57
27	3 -J	3.86	4.12	4.16 (0.03)	4.27	4.11
28	2 -J	2.86	3.49	3.51 (0.08)	3.59	3.25
29	4 -J	3.93	4.25	4.31 (0.04)	4.29	4.00
30	3 $-OCH_3$	4.09	4.00	4.00 (0.11)	4.36	4.79
31	2 $-OCH_3$	4.09	4.20	3.98 (0.17)	4.06	3.30
32	4 $-OCH_3$	4.47	4.60	4.60 (0.11)	4.19	4.60
33	3 $-CH_3$	4.27	4.35	4.37 (0.03)	4.44	4.56
34	2 $-CH_3$	3.92	4.01	3.90 (0.15)	4.00	3.86
35	4 $-CH_3$	4.36	4.39	4.44 (0.05)	4.42	4.49
36	3 $-NO_2$	3.45	2.89	2.91 (0.09)	3.13	3.50
37	2 $-NO_2$	2.17	1.49	1.74 (0.35)	2.16	2.04
38	4 $-NO_2$	3.44	3.22	3.30 (0.07)	3.34	3.78
39	2 −OH, 4 −OH, 6 −OH	1.68	1.59	1.67 (0.12)	0.91	1.52
40	2 $-NO_2$, 4 $-NO_2$, 6 $-NO_2$	0.65	0.72	0.77 (0.29)	2.42	0.44
41	H	4.18	4.25	4.20 (0.08)	4.42	4.10

[a] For the most predictive model: benzoic template, D = 80; MD = 0.8; $r^2 = 0.90$, $s = 0.32$. [b] Benzoic template, the mean predicted value for the experiments with all tested, i.e., 16 different values of the D and MD parameters, (SD) − standard deviation. [c] For the most predictive model: formic template, D = 60; MD = 0.8; $r^2 = 0.80$, $s = 0.45$. [d] CoMFA model: $q^2 = 0.75$; $s = 0.48$.

mization technique. The energy gradient convergence criterion was set to 0.05 kcal/mol. Partial charges were calculated with the empirical procedure of Gasteiger−Marsilli. The default distant-dependent dielectric model was used for the electrostatic potential calculations. The steric and electrostatic field energies were calculated using carbon sp^3 probe atom with a charge of +1. The CoMFA grid spacing was 2.0 Å for all the dimensions within the defined region, which extended beyond the van der Waals envelopes of all molecules by at least 4.0 Å. Other calculations were performed with hydrogen as a probe atom with a charge of +1. For each molecule the energies with a total of 720 grid points were calculated with 2 Å spacing in a lattice of 14 × 16 × 18. All steric energies with a value greater than 4.0 kcal/mol were truncated to 4.0. The lattice points with the deviation of less than 0.05 kcal/mol were discarded.

Data Analysis. Kohonen Mapping. The competitive Kohonen strategy[26] was used to construct a two-dimensional topographic map obtaining the signals from the points sampled randomly at the molecular surface. As molecular surfaces are continuous the plane of projection was also selected to be a continuous surface. Thus we used a torus for this purpose, which was cut along two perpendicular lines and then spread into a plane. Each neuron, j, was then defined by three weights, w_{ji}. The competitive training of the network was based on the rule that each point, s, of the molecular surface was projected into that neuron, sc, that has weights, w_{ci}, that come closest to the Cartesian coordinates, x_{si}, of this point, s (eq 1).

$$out_{sc} \leftarrow \min\Big[\sum_{i=1}^{m}(x_{si} - w_{ji})^2\Big] \tag{1}$$

A projection of the electrostatic potential value (MEP) from the surface points, s, into such a two-dimensional arrangement of neurons, after calculating the average MEP value within this particular neuron and scaling this values into the respective colors results in so-called feature map.

Comparative Kohonen Mapping. In fact, such a map illustrates the property (MEP) of a single molecule. As however, the weights of the Kohonen network contain the shape of the certain molecular surface, it can be used to compare the geometries of molecular surfaces of other

Table 2. Comparison of the Measured and Predicted pK_a Values for Alkanoic Acids **42−88** and **42−74**, Respectively

no.	acid	p$K_{(exp)}$	p$K_{(pred)}$ (SD)[a]	p$K_{(pred)}$ (SD)[b]	p$K_{(pred)}$[c]	p$K_{(pred)}$[d]
42	CH_3COCH_2COOH	3.58	3.25 (0.23)	3.72 (0.20)	2.42	3.60
43	$BrCH_2COOH$	2.90	3.52 (0.05)	3.03 (0.22)	3.37	3.23
44	$BrCH_2CH_2COOH$	4.02	4.06 (0.07)	4.30 (0.05)	4.32	4.17
45	$ClCH_2COOH$	2.83	3.25 (0.01)	3.04 (0.05)	3.15	3.04
46	CH_2ClCH_2COOH	4.08	3.96 (0.07)	4.21 (0.06)	4.27	4.18
47	$(CN)CH_2COOH$	2.47	2.78 (0.04)	2.84 (0.06)	2.56	2.29
48	$(CN)CH_2CH_2COOH$	4.44	3.65 (0.09)	3.90 (0.05)	4.20	4.24
49	$C_6H_5CH_2CH_2COOH$	4.66	4.55 (0.03)	4.74 (0.06)	4.64	4.81
50	$3\text{-F}-C_6H_4-O-CH_2COOH$	3.08	3.41 (0.02)	3.14 (0.08)	2.87	3.07
51	$2\text{-F}-C_6H_4-O-CH_2COOH$	3.08	3.49 (0.02)	3.13 (0.04)	3.57	3.10
52	$4\text{-F}-C_6H_4-O-CH_2COOH$	3.13	3.39 (0.03)	3.14 (0.08)	3.75	3.28
53	$HOCH_2COOH$	3.85	3.68 (0.02)	3.40 (0.07)	3.15	3.73
54	$C_6H_5CONHCH_2COOH$	3.64	4.08 (0.04)	3.68 (0.26)	3.32	3.62
55	$CH_3CH(OH)CH_2COOH$	4.70	3.91 (0.13)	4.34 (0.19)	4.79	4.65
56	$HOCH_2CH_2CH_2COOH$	4.72	4.39 (0.04)	4.46 (0.08)	4.72	4.81
57	$CH_2(OH)CH_2COOH$	4.51	4.24 (0.05)	4.46 (0.13)	4.36	4.48
58	$(CH_3)_2CHCH_2CH_2COOH$	4.85	4.68 (0.03)	4.78 (0.03)	3.90	4.80
59	$(CH_3)_2CHCH_2COOH$	4.78	5.03 (0.17)	4.75 (0.16)	5.30	5.05
60	JCH_2COOH	3.17	4.09 (0.06)	3.58 (0.24)	3.33	3.23
61	$CH_3CH_2CH_2CH_2CH_2COOH$	4.85	4.67 (0.03)	4.81 (0.03)	4.98	4.81
62	$CH_3CH_2CH_2CH_2CH_2CH_2CH_2COOH$	4.89	4.67 (0.03)	4.80 (0.03)	5.00	4.82
63	$CH_3CH_2CH_2COOH$	4.82	4.68 (0.03)	4.82 (0.03)	4.72	4.77
64	NO_2CH_2COOH	2.26	2.45 (0.03)	2.09 (0.12)	2.29	2.05
65	CH_3CH_2COOH	4.87	4.69 (0.03)	4.83 (0.03)	4.38	4.79
66	$HSCH_2COOH$	3.68	4.24 (0.03)	3.93 (0.15)	4.07	3.69
67	$CF_3CH_2CH_2COOH$	4.16	3.78 (0.09)	4.07 (0.07)	2.61	4.06
68	CF_3CH_2COOH	3.02	2.89 (0.03)	2.65 (0.05)	3.29	3.10
69	$CH_3CH_2CH_2CH_2COOH$	4.86	4.67 (0.03)	4.81 (0.03)	4.98	4.83
70	$CH_2{=}CHCH_2COOH$	4.34	4.56 (0.02)	4.53 (0.04)	3.92	4.21
71	$3\text{-}OCH_3-C_6H_4CH_2CH_2COOH$	4.65	4.61 (0.03)	4.82 (0.05)	4.78	4.64
72	$2\text{-}OCH_3-C_6H_4CH_2CH_2COOH$	4.80	5.15 (0.04)	4.64 (0.18)	4.15	4.79
73	$4\text{-}OCH_3-C_6H_4CH_2CH_2COOH$	4.69	4.46 (0.04)	4.69 (0.05)	5.05	4.59
74	CH_3COOH	4.75	4.68 (0.04)	4.84 (0.05)	3.93	4.60
75	$CH_3CHBrCOOH$	2.97	3.59 (0.07)	*e*	3.72	*e*
76	$CHCl_2COOH$	1.48	2.27 (0.07)		3.16	
77	$CH_3CHClCOOH$	2.85	3.47 (0.10)		3.57	
78	$(C_2H_5)_2CHCOOH$	4.73	5.00 (0.16)		5.20	
79	$CH_3CH(CN)COOH$	3.99	2.66 (0.06)		1.91	
80	$(C_6H_5)_2CHCOOH$	3.94	4.54 (0.16)		4.55	
81	$CH_3CH(C_6H_5)COOH$	4.64	4.40 (0.09)		4.76	
82	$(CH_3)_2CHCOOH$	4.85	4.72 (0.03)		3.88	
83	$C_6H_5CH(OH)COOH$	3.85	3.69 (0.06)		4.01	
84	$CH_3\ CH(OH)COOH$	3.86	3.74 (0.03)		2.87	
85	$CH_3CH_2CH(NO_2)COOH$	2.39	2.59 (0.06)		3.63	
86	CCl_3COOH	1.66	1.15 (0.07)		3.08	
87	CF_3COOH	0.23	0.24 (0.18)		1.72	
88	$C(CH_3)_3COOH$	5.05	4.75 (0.02)		3.83	

[a] Formic template, cross-validated model excluding all empty neurons; the mean predictive value for the experiments with all tested, i.e., 14 different values of the D and MD parameters, (SD) − standard deviation. [b] Formic template, cross-validated model excluding all empty neurons; the mean predictive value for experiments with all tested, i.e., 14 different values of the D and MD parameters. [c] CoMFA: $q^2 = 0.52$; $s = 0.77$. [d] CoMFA: $q^2 = 0.61$; $s = 0.55$. [e] Not included into the model.

molecules. In such a method the trained Kohonen network is processing the signals coming from the surface of other molecule(s), i.e., the electrostatic potential of each input vector was projected through the network to obtain a series of comparative maps both for the template molecule and each analyzed molecule. The respective electrostatic potential values from the surfaces of the processed molecules were then projected into such a network allowing us to compare these parts of the molecule surfaces that can be superimposed. If the surfaces cannot be superimposed on the reference molecule (template), then the respective output neurons get no signal from the molecules processed.

All the molecules were superimposed before the calculation of the molecular surfaces. The superimposition was performed by covering the following: (1) for molecules (**1−41**) all non-hydrogen atoms of carboxylic function (formic template) or all non-hydrogen atoms of benzoic acid (benzoic template) and (2) for molecules (**42−88**) all non-hydrogen atoms of carboxylic function (formic or acetic template).

We used Match3D program[27] for performing this operation. The KMAP 3.0 program[27] was used for the simulation of Kohonen networks. The size of the Kohonen networks amounts to 20 × 20 neurons. The output of this program was used for the calculation of the mean electrostatic potential values within each neuron, and respective feature maps were transformed to a respective 400 element vectors.

PLS Analysis. Vectors obtained were processed by the PLS analysis with a leave-one-out cross-validation procedure. The PLS procedures were programmed within the MATLAB environment (MATLAB).

A PLS model was constructed for the centered data, and its complexity was estimated based on the leave-one-out cross-validation procedure (CV). In the leave-one-out CV one repeats the calibration *m* times, each time treating the *i*th left-out object as the prediction object. The dependent variable for each left-out object is calculated based on the model with one, two, three, etc. factors. The Root Mean Square Error of CV for the model with *j* factors is defined as

$$\mathrm{RMSECV}_j = \sqrt{\frac{\sum_i (\mathrm{obs}_i - \mathrm{pred}_{i,j})^2}{\mathrm{m}}} \qquad (2)$$

where obs denotes the assayed value, pred denotes the predicted value of the dependent variable, and *i* refers to the object index, which ranges from 1 to *m*. Model with *k* factors, for which RMSECV reaches a minimum, is considered as an optimal one.

We used the performance metrics that are accepted and widely used in CoMFA analyses, i.e., cross-validated q^2_{cv}

$$q^2_{\mathrm{cv}} = 1 - \frac{\sum (\mathrm{obs}_i - \mathrm{pred}_i)^2}{\sum (\mathrm{obs}_i - \mathrm{mean(obs)})^2} \qquad (3)$$

where obs denotes the assayed values, pred denotes the predicted values, mean denotes the mean value of obs, and *i* refers to the object index, which ranges from 1 to *m*, and cross-validated standard error *s*

$$s = \sqrt{\frac{\sum (\mathrm{obs}_i - \mathrm{pred}_i)^2}{m - k - 1}} \qquad (4)$$

where *m* is the number of objects, and *k* is the number of PLS factors in the model.

Before PLS analysis was performed the descriptors were centered, and this operation was repeated for each cross-validation run.

RESULTS AND DISCUSSION

Parameters Controlling the Comparative Kohonen Mapping. The Kohonen procedure was used both to project the MEP of the single molecule—to form a single topographic Kohonen map—or the series of the molecules—to give a series of the comparative maps. These techniques have been applied to investigate and describe molecular effects. The advantages of both methods have been discussed in previous publications.[16] Although the method with a single molecule procedure proved its usefulness, there are however some complications. For example, the dependence of the final pattern upon many parameters controlling the performance of neural network, and consequently the dependence upon the software used for simulation, seems to be ones of the most important. In contrast, we have observed that a use of comparative technique can provide highly reproducible results irrespective of the software used, in particular, while comparing the KMAP program and the Kohonen SOM toolbox.[28,29] Shown below is that this technique gives not only reproducible but also very stable results.

For a better understanding Figure 1 gives a schematic view illustrating the operation realized by the Kohonen network while comparing the surfaces of two molecules, i.e, a template molecule TM and a molecule M1. First, the operation performed by the network can be interpreted as the clustering of the surface points found within the certain winning distance (or maximal distance “MD”) into the single output neuron. As the template data was used for training of the template network the spheres defining winning diameter, e.g. A1, A2, A3, were based on the surface of the template molecule. Therefore, for two distinct surfaces two main possibilities can occur. Either a winning distance is large enough to include some points on the M1 surface (MD1, MD2) or it does not include any points on the M1 surface (MD3). In the latter case the empty neurons appeared within the comparative pattern. Therefore, it seems to us that the radius of MD that is to be defined during computations should be the most important parameters controlling the efficiency of the comparison. The second important parameter is the density of the points sampled at both surfaces (D).

The pK_a Values of o-, m-, and p-Substituted Benzoic Acids. The Hammet equation is probably the best known quantitative relation between the structure and reactivity of benzoic acids. Hammet constant gives an account of electronic effect both of the inductive and resonance background. Due to the steric interference, o-analogues are not included into the Hammet model. The Hammet constant was modeled previously using 3D CoMFA[30,31] and CoMSA[32] techniques. However, we have not found any 3D QSAR study that includes p- and m-analogues together with o-analogues. Below we describe the attempt to predict the pK_a values for o-, m-, and p-substituted analogues of benzoic acid.

We chose two templates to display two different regions of the series. The first one is benzoic acid (we will indicate such a template as a benzoic template) that allows analyzing the region of the benzene ring and carboxylic function together. The second template is formic acid (we will indicate such a template as a formic template), which visualizes the carboxylic region only. Figure 2 illustrates some of the patterns given during such a procedure for the benzoic template.

We performed a complex study analyzing both the neutral and anionic form of the carboxylic function. We found that both the anionic and acidic form of the molecules provide comparable results. The results presented below are representative for both cases. The performance of some of the PLS analysis of the comparative MEP matrices are compared in Table 3 and Figures 3−6, respectively. The first series of the analyses was performed while optimizing the number of the PLS components included into a final model. Although we performed such analyses for all examples discussed, we illustrate this particular case only for the examples shown in Figure 3. The components included into the respective models were extracted from the six latent PLS components, which means the maximal number of the component included in the final model can amount to 6. The optimal number of components range from 3 to 6. Figure 3a,b analyzes the quality of the models resulted for the anionic forms projected on the benzoic template (the optimal number of components 3−5). The analysis involving acidic form, which is not shown in Figure 3, provides only slightly different results. The optimal values (acidic form) amount to $q^2_{\mathrm{cv}} = 0.90$ and $s_{\mathrm{cv}} = 0.32$ with six components (benzoic template; the MS1b model from Table 3) vs $q^2_{\mathrm{cv}} = 0.80$ and $s_{\mathrm{cv}} = 0.45$ with six components (formic template; the MS1a model from Table 3). Thus, the most important conclusion from this part is

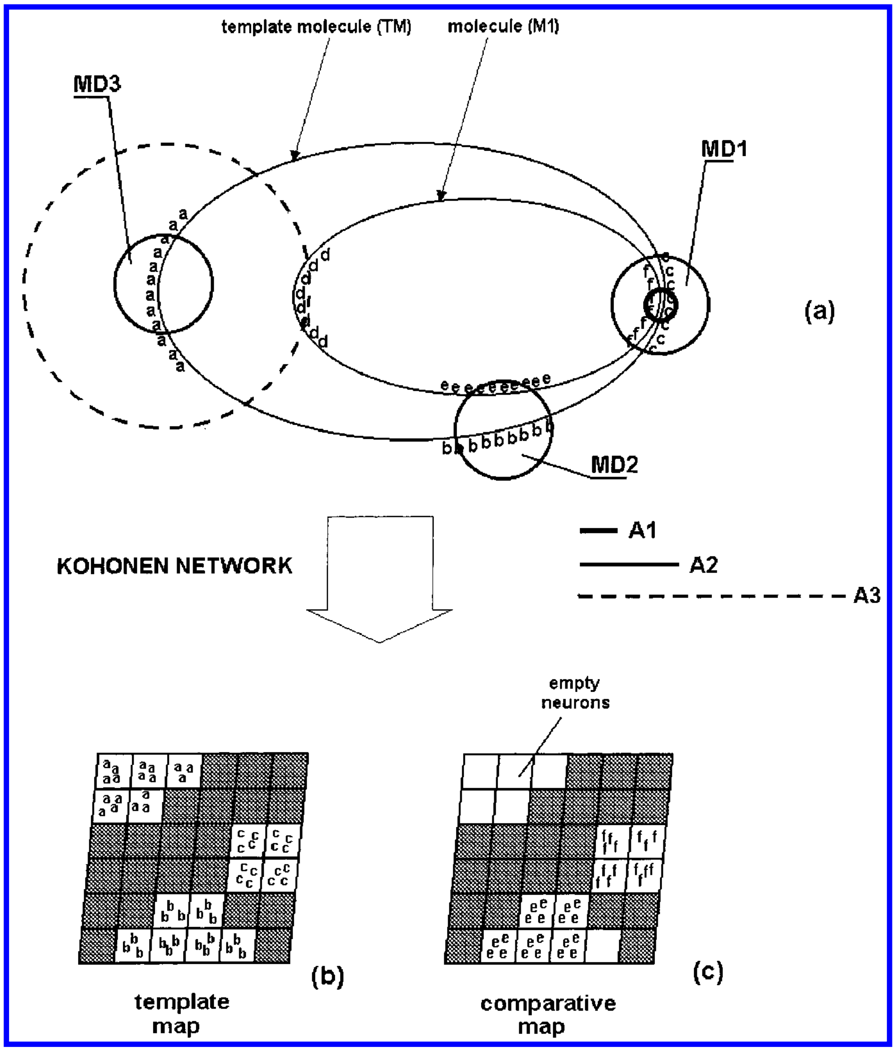

Figure 1. Schematic view illustrating the operation realized by the Kohonen network while comparing the surface of two molecules, i.e., a template molecule TM and a molecule M1 (a). Kohonen neural network operates by grouping in single neurons all the input points found within an environment of a certain diameter MD1, MD2, and MD3 (maximal distance or winning distance) to give a template map (b) including only very few empty neurons. Therefore, the comparative map (c) can include the nonempty neuron if the surface points of the counter-template can be found within winning distance; (MD1, MD2) or the empty neuron if none of the surface points are present in the winning distance sphere (MD3). The comparative map (c) was constructed for the MD parameters defined by the A2 diameter.

that generally the benzoic template provides better results (higher q^2_{cv} and lower s_{cv} values) than the formic template. These results are opposite to those obtained in previous CoMFA and CoMSA analyses for p- and m-substituted analogues, which have indicated that better models can be obtained while analyzing the carboxylic region alone.[32] In this particular case, however, steric effects influencing the pK_a values of the o-analogues included in our analysis can explain the difference. As only the benzoic template can provide the patterns visualizing the influence of steric effects, it provides better models than the formic template.

To evaluate the influence of the D and MD parameters we performed the second series of calculations keeping a common number of components in all models (Figure 3c,d and 4). Generally the influence of the D and MD parameters is far less important than was expected. One can try to anticipate the influence of the D and MD by the analysis of the importance of the steric and electrostatic effects which both are needed to explain pK_a values of the combined o- with m- and p-analogues. Steric effects can be observed within the comparative maps indirectly, mainly as differences between various conformational types. They are manifested by different patterns of empty neurons, as those observed in Figure 2. A lower MD value, that provides more rigorous shape comparison, should allow consequently for the better account of steric effects, providing more predictive models. On the other hand, the lower density D decreases the number of points to be processed by the network during training. Consequently, this operation can be more optimal, giving the map a slightly better quality, with regards to the minor differences of the empty neuron pattern. This also implies a more accurate geometrical comparison.

Actually, for the benzoic template we can observe that practically both parameters do not influence the performance of the network (Figure 3). As the formic template discloses only electrostatic effects, the rigidity of the shape comparison is not so important. The regularity observed indicates that it is rather the area of the surface from which the points are attracted into an individual neuron that decides the quality of the model. Therefore, higher values of the MD parameters, favoring a larger area, also increase the quality of final model (Figure 4).

To further prove these conclusions we attempted to mask any possible influence of steric effects displayed within the

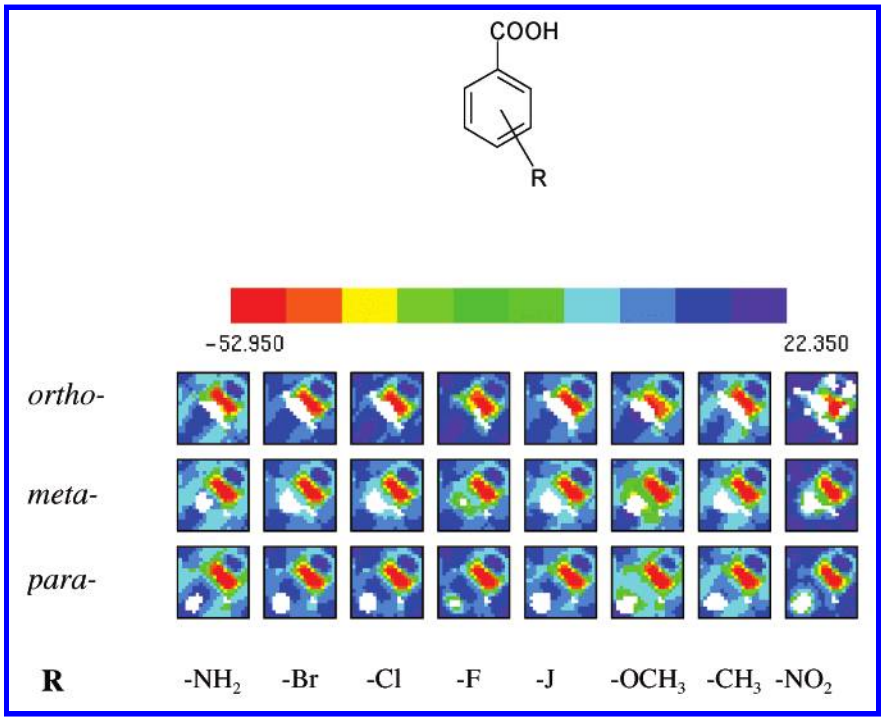

Figure 2. The comparative Kohonen patterns obtained during mapping some o-, m-, and p-benzoic acids (benzoic template). The white motif of the empty neurons strictly defines each series, indicating the difference of the respective benzene ring substituent and the hydrogen atom of the benzoic acid template.

Table 3. Comparison between CoMFA and CoMSA Models, Respectively

	CoMFA[a]				CoMSA[b]			
compounds	model	q^2	s	template	model	q^2	s	template
benzoic acids **1**−**41**	MF1	0.75	0.48	benzoic acid	MS1a	0.80	0.45	formic acid
					MS1b	0.90	0.32	benzoic acid
alkanoic acids **42**−**74**	MF2	0.61	0.55	formic acid	MS2a	0.94	0.21	formic acid
					MS2b	0.92	0.26	acetic acid
alkanoic acids **42**−**88**	MF3	0.52	0.77	formic acid	MS3a	0.86	0.42	formic acid
					MS3b	0.82	0.49	acetic acid

[a] Hydrogen atom probe; the acidic form of the molecules were used to construct the models reported. [b] The acidic form of the molecules were used to construct the models reported.

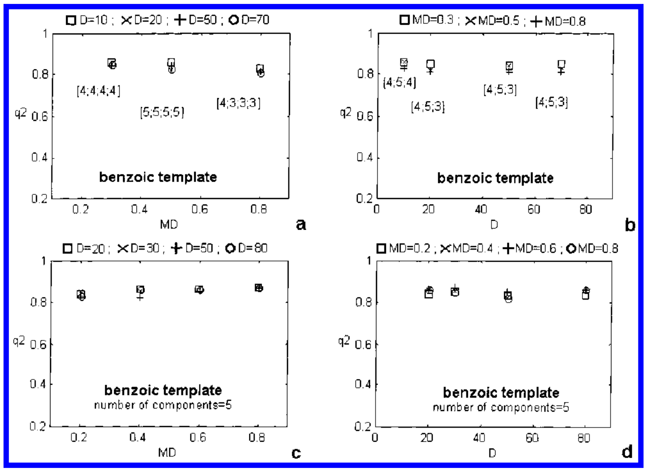

Figure 3. The relationship between the q^2_{cv} performance of the CoMSA models and MD or D parameters, respectively, for anionic (a,b) and acidic (c,d) form of the acid molecule. Bracketed numbers (a,b) indicate the optimal number of the PLS components.

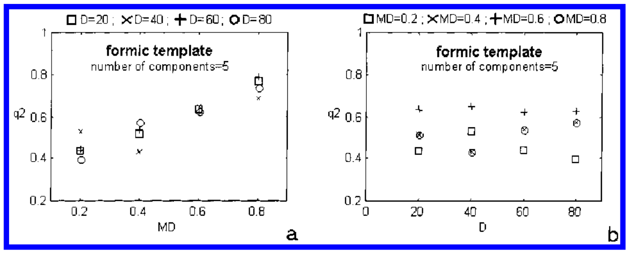

Figure 4. The relationship between the q^2_{cv} performance of the CoMSA models and MD or D parameters, respectively, for acidic (a, b) form o-, m-, and p-analogues of the benzoic acids.

map. This was achieved by including into the analysis only these neurons of the maps that are nonempty within all the maps (nonempty maps "nen"). Due to the need of description of steric effects, it can be speculated that the "nen" results should be worse than those including all neurons. In fact, as anticipated, the PLS analysis of the full maps for both templates provides more predictive models than this of the *nen* maps (Figures 5 and 6). Further practical conclusion from this part is that formic template cannot be a proper choice for the series analyzed. The exclusion of steric description makes the models much worse and unstable, e.g. as the ones shown in Figure 6.

The analysis discussed above allowed us to design an optimal condition (benzoic template, MD = 0.8, D = 80 points/Å^2) for quantitative prediction of the pK_a values. Table 1 (column 4) compares the cross-validated pK_a values (leave-

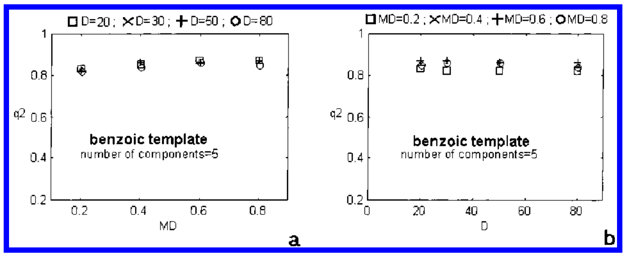

Figure 5. The relationship between the q^2_{cv} performance of the "nen"-CoMSA (nonempty neurons only—details in text) models and the MD or D parameters, respectively, for acidic (a, b) form o-, m-, and p-analogues of benzoic acids.

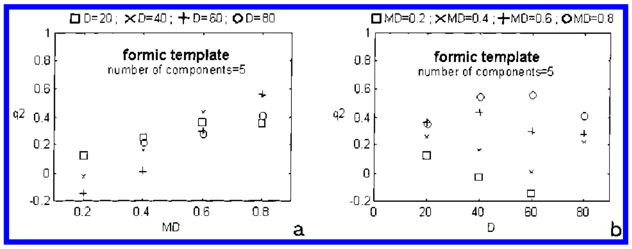

Figure 6. The relationship between the q^2_{cv} performance of the "nen"-CoMSA (nonempty neurons only − details in text) models and the MD or D parameters, respectively, for acidic (a, b) form o-, m-, and p-analogues of benzoic acids.

one out procedure) with the actual ones (column 3). In addition, column 5 shows the average pK_a values obtained in the series of simulations including all tested D and MD parameters, i.e., for 16 different D and MD values. It can be observed that these values only slightly differ from the optimal ones.

To compare the performance of the CoMSA method with standard procedures we also performed a typical CoMFA analysis for the series. In this particular case the neutral form of the acid provided slightly better results. The best model was obtained for CoMFA with the H^+ probe atom. This is given by $q^2_{cv} = 0.75$ and $s_{cv} = 0.48$ (model MF1 from Table 3), which means in this particular case CoMFA is less effective than CoMSA. On the other hand it is worth noting that these values correspond quite well with CoMSA using formic template ($q^2_{cv} = 0.80$; $s_{cv} = 0.45$; model MS1a from Table 3). This conclusion clearly indicates that electronic effects predominate in CoMSA/(formic template) and CoMFA/H^+.

The pK_a Values of the Alkanoic Acids. A change of the pK_a values of the analogues of alkanoic acids having different substituents in the alkyl chain is one of the substantial effects described and discussed in the handbooks of organic chemistry. Electronic effects determine the pK_a values within such a series.

Table 2 shows the results for the alkanoic acids analyzed. Analogues **42**−**74** are the acids having two hydrogen atoms at the C2 carbon atom. We used the formic and acetic templates to display the maps of this series. Table 2 compares the actual and predicted pK_a values of the acids. Like previously, we tested the stability of the models by performing a series of experiments, which were analyzed by the average values and standard derivation. The best model obtained for a formic template is described by the $q^2_{cv} = 0.94$ and $s_{cv} = 0.21$ (MS2a model from Table 3). Figures 7 and 8 provide the analysis of the influence of the D and MD parameters upon the predictability of the models including acids **42**−**74**. It can be observed that the results are quite stable and the influence of the D and MD is rather weak. As can be anticipated, now the exclusion of empty neurons—the *nen* analysis—slightly increases the quality of the respective models.

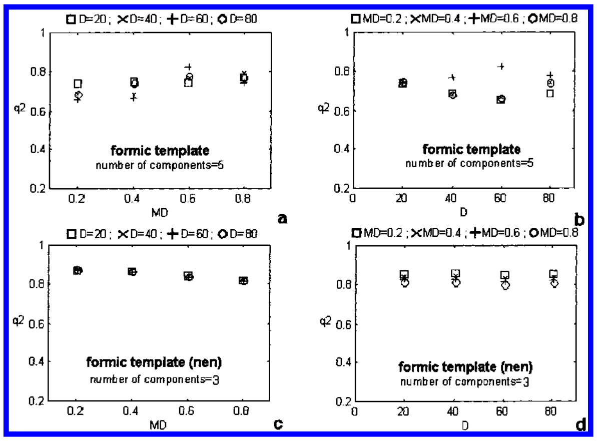

Figure 7. The relationship between the q^2_{cv} performance of the CoMSA (a,b) and "nen"-CoMSA (nonempty neurons only − details in text) (c, d) models and the MD or D parameters, respectively, for the acidic form of the alkanoic acids **42**−**74**.

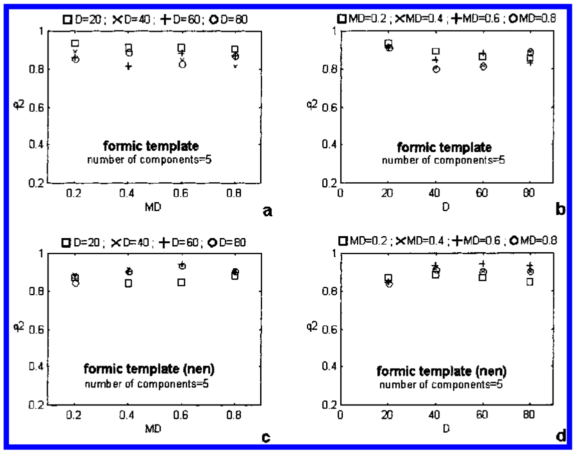

Figure 8. The relationship between the q^2_{cv} performance of the CoMSA (a, b) and "nen"-CoMSA (nonempty neurons only − details in text) (c, d) models and the MD or D parameters, respectively, for the acidic form of the alkanoic acids **42**−**88**.

Table 3 gives also the results for the similar experiment involving a broader range of alkanoic acids **42**−**88**. The predictivity of the model is now slightly lower, amounting to $q^2_{cv} = 0.86$ and $s_{cv} = 0.42$ (MS3a model from Table 3), but the results are very stable. Extending the area of the template from formic to acetic acid provides models of the slightly lower predictivity (the MS2b and MS3b models from Table 3). The comparison of the CoMSA and CoMFA models (Table 3) indicates CoMSA, as previously**,** provides better description of the series.

CONCLUSIONS

A self-organizing neural network was used to design a novel method capable of the quantitative prediction of molecular properties. The method is based on the comparison of molecular surfaces performed by the coupled neural network and PLS system. Unlike CoMFA and related methods it compares the properties describing not a discrete set of points but the average property value calculated for a certain area of the molecular surface. It has been found that the results of the PLS analysis of the series of the comparative matrices of the molecular electrostatic potential (MEP) are quite stable and only slightly depends on such parameters as the number of points sampled at the molecular surface (D) or a winning distance (MD) of the self-organizing neurons. The influence of these parameters for modeling the effects limited by steric and electronic effects was determined, and the pK_a values of the o-, m-, and p-analogues of benzoic acid and alkanoic acid were predicted.

ACKNOWLEDGMENT

The authors thank Professor Johann Gasteiger of the University of Erlangen-Nürnberg, BRD both for his valuable discussion and for facilitating the programs of CORINA, PETRA, SURFACE, and KMAP. The financial support of the KBN Warsaw, grant numbers PBZ KBN - 040 P04/08 and T08E02820, is gratefully acknowledged.

REFERENCES AND NOTES

(1) Kubinyi, H. QSAR and 3D QSAR in drug design. Part 1: Methodology. *Drug Discovery Today* **1997**, *2*, 457−467.

(2) Kubinyi, H. QSAR and 3D QSAR in drug design. Part 2: Applications and problems. *Drug Discovery Today* **1997**, *2*, 538−546.

(3) Kubinyi, H. QSAR: Hansch analysis and related approaches. In *Methods and principles in medicinal chemistry*; Mannhold, R., Krokgsgaard-Larsen, P., Timmerman, H., Eds.; VCH: Weinheim, 1993.

(4) Katrizky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Structurally diverse quantitative structure − property relationship correlations of technologically relevant physical properties. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1−18

(5) Gao, H.; Katzenellenbogen, J. A.; Hansch, C. Comparative QSAR analysis of estrogen receptor ligands. *Chem. Rev.* **1999**, *99*, 723−744.

(6) Garg, R.; Gupta, S. P.; Gao, H.; Babu, M. S.; Debnath, A. K.; Hansch, C. Comparative quantitative structure−activity relationship studies on anti-HIV drugs. *Chem. Rev.* **1999**, *99*, 3525−3601.

(7) Kurup, A.; Garg, R.; Hansch, C. Comparative QSAR analysis of 5α-reductase inhibitors. *Chem. Rev.* **2000**, *100*, 909−924.

(8) Kim, K. H.; Greco, G.; Novellino, E. A Critical review of the recent CoMFA applications. *Perspect. Drug Discov. Design* **1998**, *12*/*13*/*14*, 257−315.

(9) Martin, Y. C. 3D QSAR: Current state, scope and limitations. *Perspect. Drug Discov. Design* **1998**, *12*/*13*/*14*, 3−23.

(10) Norinder, U. Recent progress in CoMFA methodology and related techniques. *Perspect. Drug Discov. Design* **1998**, *12*/*13*/*14*, 25−39.

(11) Kroemer, R. T.; Hecht, P.; Guessregen, S.; Liedl, K. R. Improving the quality of CoMFA models. *Perspect. Drug Discov. Design* **1998**, *12*/*13*/*14*, 41−56.

(12) Jain, A. N.; Koile, K.; Chapman, D. Compass − predicting biological activities from molecular-surface properties − performance comparisons on a steroid benchmark. *J. Med. Chem.* **1994**, *37*, 2315.

(13) Manallack, D. T.; Livingstone, D. J. Neural networks in drug discovery − have they lived up to their promise. *Eur. J. Med. Chem.* **1999**, *34*, 195−208.

(14) Robinson, D. D.; Winn, P. J.; Lyne, P. D.; Richards, W. G. Self-organizing molecular field analysis: A tool for structure−activity studies. *J. Med. Chem.* **1999**, *42*, 573−583.

(15) Anzali, S.; Gasteiger, J.; Holzgrabe, U.; Polanski, J.; Teckentrup, A.; Wagener, M. The use of self-organizing neural networks in drug design. *Perspect. Drug Discov. Design* **1998**, *9*/*10*/*11*, 273−299.

(16) Zupan, J.; Gasteiger, J. *Neural Networks and drug design for Chemists*, 2nd ed.; VCH: Weinheim, 1999.

(17) Gasteiger, J.; Li, X.; Rudolph, Ch.; Sadowski, J.; Zupan, J. The representation of molecular electrostatic potentials by topological feature maps. *J. Am. Chem. Soc.* **1994**, *116*, 4608−4620.

(18) Polański, J. The receptor-like neural network for modeling corticosteroid and testosterone binding globulins. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 478−484.

(19) Anzali, S.; Barnickel, G.; Krug, M.; Sadowski, J.; Wagener, M.; Gasteiger, J.; Polański, J. The comparison of geometric and electronic properties of molecular surfaces by neural networks: Application to the analysis of corticosteroid globulin activity of steroids. *J. Comput.-Aided Mol. Design* **1996**, *10*, 521.

(20) Physical and Chemical Data Compendium; Poradnik fizykochemiczny, WNT: Warsaw, 1974; pp 347−351.

(21) Gasteiger, J. CORINA for the information, see: http://www.mol-net.de.

(22) Sadowski, J.; Gasteiger, J. From atoms and bonds to three-dimensional atomic coordinates: Automatic model builders. *Chem. Rev.* **1993**, *93*, 2567−2581.

(23) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000−1008.

(24) Gastaiger, J.; Saller, H. Calculation of the charge distribution in conjugated systems by a quantification of the resonance concept. *Angew. Chem.* **1985**, *97*, 699−701.

(25) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity − a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.

(26) Kohonen, T. Self-organization and associative memory, 3rd ed.; Springer: Berlin, 1989.

(27) Gasteiger, J. Match3D; KMAP for the information, see: http://www2.ccc.uni-erlangen.de.

(28) Polański, J. The mapping of the molecular surfaces by means of self-organizing neural networks within MATLAB 5.2 for WINDOWS-95. *Acta Pol. Pharm.* **1999**, *56*, 80−84.

(29) Kohonen, T. freeware available at http://www.cis.hut.fi/resrearch/reports/quinquennial/ch4.pdf.

(30) Kim, K. H.; Martin, Y. C. Direct prediction of linear free energy substituent effects from 3D structures using comparative molecular field analysis. 1. Electronic effects of substitued benzoic acids. *J. Org. Chem.* **1991**, *56*, 2723−2729.

(31) Martin, Y. C.; Lin, T. C.; Hetti, Ch.; DeLazzer, J. PLS Analysis of distance matrices to detect nonlinear relationships between biological potency and molecular properties. *J. Med. Chem.* **1995**, *38*, 3009−3015.

(32) Polanski, J.; Walczak, B. The comparative molecular surface analysis (CoMSA): a Novel tool for molecular design. *Comput. Chem.* **2000**, *24*, 615−625.

CI010031T