Molecules **2004**, *9*, 1148–1159

molecules

ISSN 1420-3049
http://www.mdpi.org

Self-organizing Neural Networks for Modeling Robust 3D and 4D QSAR: Application to Dihydrofolate Reductase Inhibitors

Jaroslaw Polanski *, Andrzej Bak, Rafal Gieleciak and Tomasz Magdziarz

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

* Author to whom correspondence should be addressed; e-mail: polanski@us.edu.pl

Received: 31 May 2004 / Accepted: 21 Oct 2004 / Published: 31 December 2004

Abstract: We have used SOM and grid 3D and 4D QSAR schemes for modeling the activity of a series of dihydrofolate reductase inhibitors. Careful analysis of the performance and external predictivities proves that this method can provide an efficient inhibition model.

Keywords: Self-organizing neural network, 3D QSAR, 4D QSAR, SOM-4D QSAR, CoMSA.

Introduction

Drug discovery is a complex issue that lacks a general approach. Drugs are mainly synthetic products developed by chemists. However, in this context *the most fundamental and lasting objective of synthesis is not a production of new compounds but the production of properties,* as commented by Hammond and cited by Sharpless and co-workers [1]. This fact clearly makes QSAR (Quantitative Structure-Activity Relationships), in its broadest sense, an essential and irreplaceable method in this field. However, more and more sophisticated tools are needed for the transformation of the molecular structure into the compound property space. Generally, the drug-receptor interactions are complex phenomena, which cannot be easily described. Therefore, a QSAR strategy of the comparison of a series of drug ligands separately from the receptor structure has evidently limitations. In 3D or 4D QSAR molecular superimposition that should be performed for the compound series can be mentioned here as an illustrative example. By performing superimposition, intentionally or by default, but generally independently from the receptor structure, we are assuming that molecular recognition proceeds *with exactly the same mechanism and in the same place within the receptor macromolecule.* The so-called similarity paradox (very similar molecules can evoke completely different biological activity) clearly proves that in reality this assumption is not true. Thus we need to make QSAR insensitive, as much as possible, to the noise that may appear in the data. The application of robust modeling methods [2], i.e., such that they are resistant to uncertain data may be a key to success. Neural networks can be an example of such a technique that has been successfully used in drug design [3-5]. On the other hand, multivariate nonlinear regression has also been reported as an efficient alternative to neural techniques [6-8].

Our previous publications described possible applications of self-organizing neural networks for modeling 3D and 4D QSAR [9-18]. A self-organizing neural network is an unsupervised learning scheme consisting only of a single layer, usually two-dimensional rectangular or hexagonal grid of nodes (neurons). Different distance metrics can be used to define neighborhood relations between these neurons [19-22]. SOM (self-organizing map) network is designed to process multidimensional (N-dimensional) data vectors by distributing them between the neurons in such a way that similar inputs are put closer (into the neurons that are closer neighbors) to each other than those less similar. It is worth mentioning that the method preserves the topology of the processed input object.

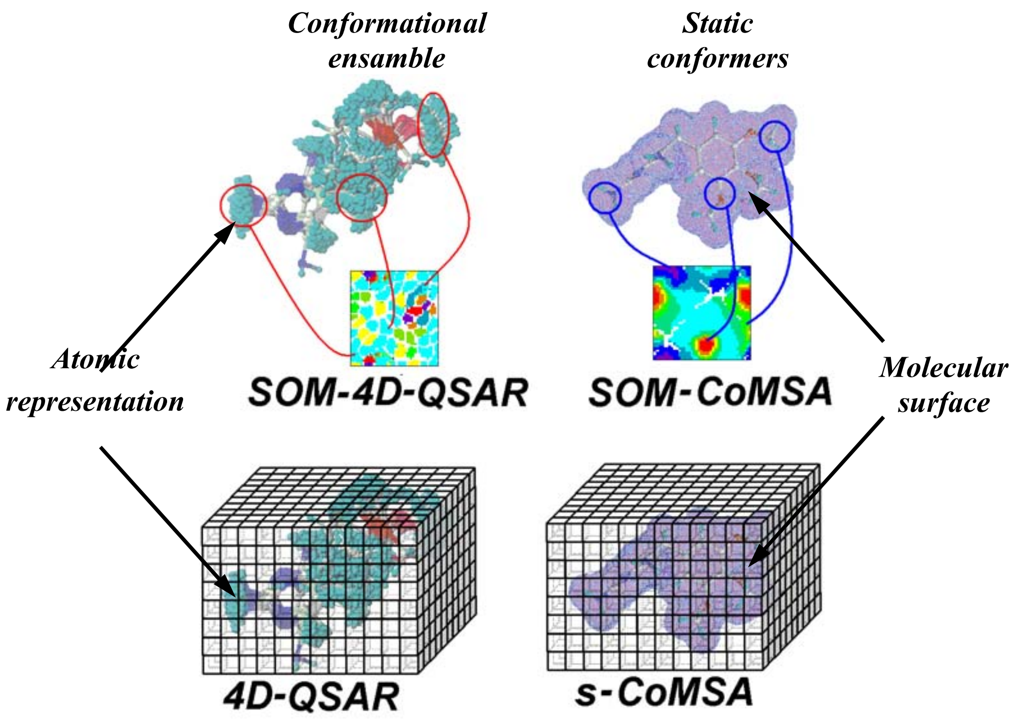
Considerable progress can be observed over the past decades in molecular development and design, in particular, in drug design. This includes new emerging disciplines and strategies that have appeared in this field. Combinatorial chemistry came as a first alternative to the traditional design and synthetic techniques. Genomics and related fields (e.g., chemo- and pharmacogenomics) have brought about an explosion of the data available for molecular design. The question may arise as to how much these new directions influence traditional methods. Do we still need traditional or multidimensional (3D or 4D) QSARs? Have traditional methods profited from these new directions? It can be clearly observed that generally, unnecessarily increasing the number of analyzed molecules, new methods investigate much larger *data populations*, irrespective of any technical problems encountered in such cases. 4D QSAR can be an example of such a technique. Basically, 4D QSAR investigates the conformational space of the molecular objects. However, in this calculation we generate for a single molecule the enormous number of conformers that investigates different spatial region. Actually, it is the likelihood of a formation of common 3D patterns of a series of molecules that is sought after by the molecular dynamics simulations. Many-fold replications of the molecules by different conformer representations allow for the increase of the chances for proper receptor structure mapping by the respective ligand structures. All this makes 4D QSAR scheme of the much more probabilistic nature, if compared to the 3D-QSAR.

In this publication we discuss the application of the SOM neural network for a QSAR scheme, in particular the SOM-4D-QSAR. Moreover, we compare this method with Comparative Molecular Surface Analysis (CoMSA) – a 3D QSAR method by the SOM neural network coupled with the Partial Least Squares Analysis (PLS) for a series of dihydrofolate reductase (DHFR) inhibitors [23].

Results and Discussion

Scheme 1 illustrates the methods used, i.e. SOM-CoMSA [9], s-CoMSA (sector – CoMSA) [24], grid-4D-QSAR and SOM-4D-QSAR [17].

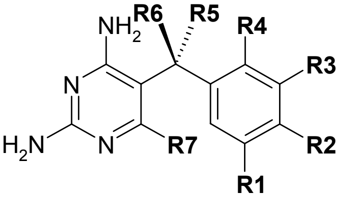
Scheme 1.



A series of dihydrofolate reductase inhibitors (DRI) are given in Table 1.

Hopfinger *et al.* analyzed this series in a publication that introduces a 4D QSAR method [23]. Thus we decided to use the same series for the comparison of the SOM versions of 3D and 4D QSAR methods. Several 3D techniques failed to model QSARs for these compounds; however; Hopfinger's 4D QSAR appeared to give a final regression equation (R^2=0.957, q^2=0.885, s=0.34) optimized by genetic algorithm (GA), performed after initial PLS. Instead of GA, we used in our method the PLS algorithm coupled with variable elimination. It is usually believed that variable elimination is not as important in PLS modeling as in standard regression procedure, because basically data transformed by PLS include this *part* of the original data that is essential for the description of the appropriate answer. However, data elimination can also be applied in PLS modeling, e.g. in Uninformative Variable Elimination (UVE) method developed by Centner *et al.* [25]. Compare references [7, 26] for the detailed investigations of variable selection in multiregression.

Table 1. The set of substituted 2,4-diamino-5-benzylpyrimidine inhibitors of *Escherichia coli* DHFR and their activity data [23].



No.	R1	R2	R3	R4	R5	R6	R7	log(1/I$_{50}$)
1	OCH_3	OCH_3	OCH_3	H	H	H	H	8.23
2	OCH_3	OCH_3	OCH_3	CH_3	H	H	H	5.85
3*R*	OCH_3	OCH_3	OCH_3	H	OH	CH_3	H	4.00
4*S*	OCH_3	OCH_3	OCH_3	H	OH	CH_3	H	4.00
5	OCH_3	OCH_3	OCH_3	H	$=CH_2$		H	5.60
6*R*	OCH_3	OCH_3	OCH_3	H	H	CH_3	H	5.35
7*S*	OCH_3	OCH_3	OCH_3	H	H	CH_3	H	5.35
8	OCH_3	Br	OCH_3	H	H	H	H	8.53
9	OCH_3	OH	OCH_3	H	H	H	H	7.96
10	OCH_3	OH	OCH_3	H	H	H	CH_3	6.52
11	OCH_3	OCH_3	OCH_3	H	H	H	CH_3	7.00
12	OH	H	OH	H	H	H	H	2.78
13	H	H	H	H	H	H	H	5.71
14	CH_2OH	H	CH_3OH	H	H	H	H	5.83
15	H	H	Cl	H	H	H	H	6.14
16	H	Br	H	H	H	H	H	6.30
17	OCH_3	H	H	H	H	H	H	6.40
18	OCH_3	H	OCH_3	H	H	H	H	7.75
19	CH_3	H	CH_3	H	H	H	H	7.45
20	H	C_6H_5	H	H	H	H	H	6.40

In our previous publications we have shown that this method as well as its modifications, i.e., modified UVE (m-UVE) and iterative variable elimination (IVE) can be used in 3D QSAR schemes [11]. This allows identifications of the molecular areas important for the interactions with biological receptors or enzymes, so-called interaction pharmacophore elements (IPEs).

Figure 1. The 4D QSAR models of the DRI series, performed using the occupancy and charge type IPEs. The numbers indicate the q^2 and s performances, and optimal number of the PLS latent variables included in the model; after UVE and (modified procedure [11]) IVE data elimination, respectively

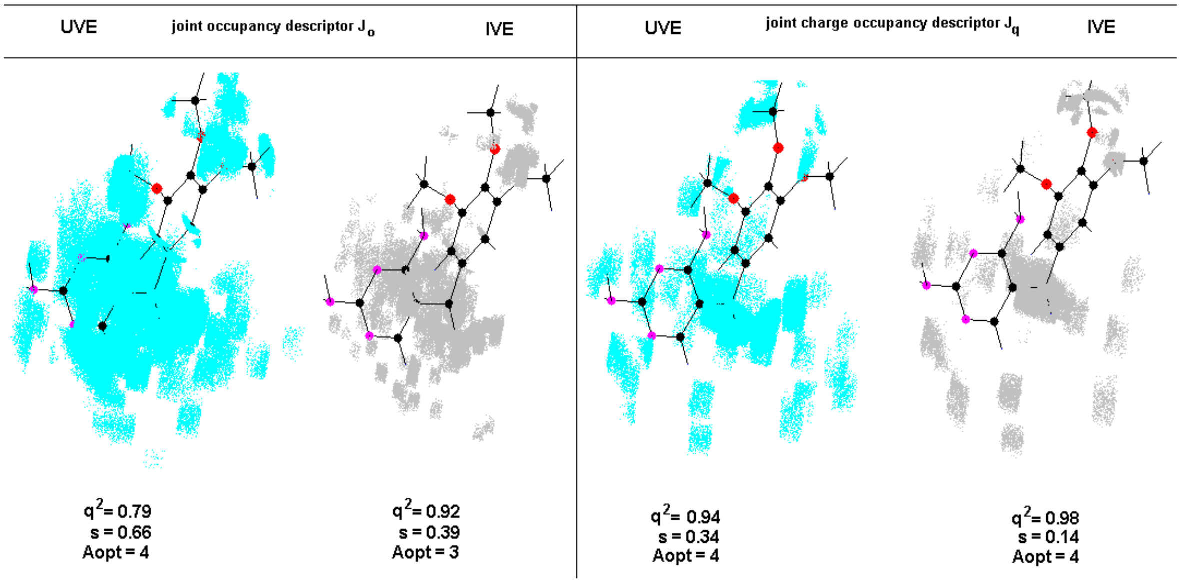
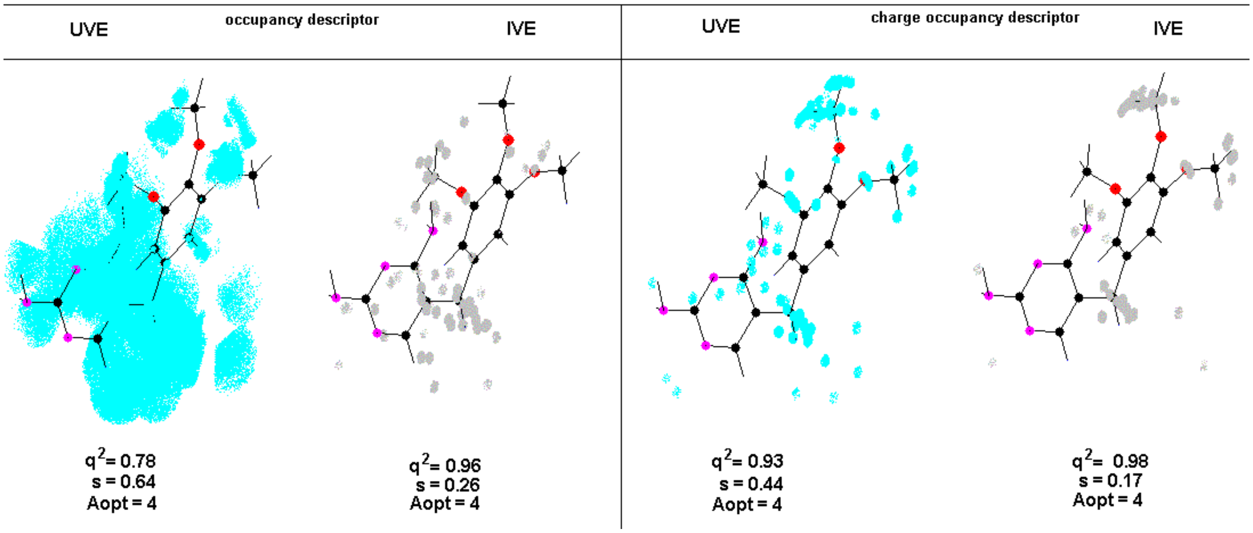


Figure 2. The SOM-4D QSAR models of the DRI series, performed using the occupancy and charge type IPEs. The numbers indicate the q^2 and s performances, and optimal number of the PLS latent variables included in the model; after UVE and IVE data elimination, respectively



The performance of the 4D QSAR PLS models obtained without data elimination ranges from q^2=0.30-0.43 and for the best model takes a value of q^2= 0.43, s=1.10, with 4 PLS components. After variable elimination these values can improve, as shown in Figures 1 and 2. This outperforms classical Hopfinger's 4D QSAR with GA. However, these data refer to the series without molecule **12**, which is an evident outlier according to our results. This can result from a fact that the activity of this compound evidently differs from the rest of the series. This may indicate some differences in the drug receptor interaction mechanism.

During modeling we always estimated an optimal number of the PLS components, but the maximal model complexity (a number of PLS components) was truncated not to exceed four. Our results only slightly depend upon the method used, i.e., classical grid method or its SOM version, and superposition mode. Figures 1 and 2 compare the IPEs revealed for DRI by SOM-4D-QSAR to that of 4D QSAR-PLS-UVE (IVE, m-UVE) methods. The performances have also been compared.

In the original publication Hopfinger *et al.* did not perform any additional model validation. However, according to the current knowledge, the q^2 value is not a sufficient criterion for verifying model quality. Thus, the description of the series using a few original variables (individual grid cells) without validation of the external predictions seems to be risky. Therefore, we divided the series into two groups of the training (compounds: 1-11 and 13-15) and test sets (compounds: 16-20) and verified model calculated for the training set by the residual error estimated for the values predicted in the test set. The best model was obtained for grid-4D-QSAR with joint occupancy type descriptors (I_c), which is characterized by q^2= 0.96, s=0.10 and standard deviation of error of prediction (SDEP) = 0.64 (with 10 PLS components. Of course a high number of PLS components makes a problem. On the other hand, this complexity is determined by model optimization. After data elimination that force lower complexity the performance is only slightly lower: q^2=0.96, s=0.12, SDEP = 0.61 with 4 PLS components. Thus, in this particular case lower complexity does not improve model predictability (SDEP value).

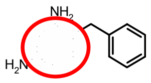
We think there are few interesting observations that appeared from the analysis of the results given in Figure 1. Thus, the molecular areas indicated by our analysis with the *occupancy type descriptors* are similar to those revealed in Hopfinger's work, in which *these type descriptors* were also used. The inclusion of *charge descriptors* improves model quality given by the q^2 and SDEP values. There is also some important regularity that can be observed during data elimination in the PLS model. In fact, we observed that for the models of the high predictivity data elimination cannot improve model quality. However, for the poor initial models, UVE (IVE) data elimination can bring an important improvement.

Both SOM- and grid- 4D-QSAR analyzed schemes provide comparable results. Table 2 compares performances of these schemes with 3D-QSAR modeling of the series activity. We used for this purpose the SOM and sector version of the Comparative Molecular Surface Analysis [24]. The results of 3D QSAR modeling are evidently worse than 4D QSAR. This indicates that conformational flexibility of the benzylpirymidine series makes 4D QSAR more efficient in modeling their inhibiting properties.

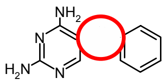
Conclusions

We used SOM and grid 3D and 4D QSAR schemes for modeling the activity of a series of dihydrofolate reductase inhibitors. We used PLS with UVE (IVE) for modeling all schemes. Careful analysis of the performances and external predictivities proves that this method can provide an efficient inhibition model.

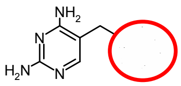
Table 2. 3D QSAR results.



a



b



c

		Superposition mode		
		a	b	c
		CoMSA		
all [b]	MD [a]	0.5	0.5	0.5
	q^2	0.62	0.72	0.64
	S	1.17	1.01	1.08
Training/test set	MD	0.5	0.5	0.5
	q^2	0.59	0.64	0.71
	S	1.02	0.91	0.90
	SDEP	1.25	0.96	1.20
IVE	max A [c]	6	5	6
	MD	0.5	0.5	0.5
	q^2	0.62	0.87	0.71
	S	0.89	0.58	0.79
	SDEP	0.81	0.72	0.78
		s-CoMSA		
all	sector size	1	1	3
	q^2	0.38	0.56	0.47
	s	1.50	0.90	1.01
Training/test set	sector size	1	1	1
	q^2	0.54	0.70	0.69
	s	1.96	1.59	1.26
	SDEP	1.42	1.32	1.42
IVE	sector size	1	1	4
	max A	4	1	3
	q^2	0.73	0.59	0.68
	s	0.83	0.92	0.85
	SDEP	0.93	1.15	1.10

[a] MD – Maximal distance for comparative Kohonen maps [10] [b] Models without variable elimination [c] Maximal number of PLS components during variable elimination procedure (IVE-PLS) [11].

Acknowledgments

The authors thank Professor Johann Gasteiger of the University of Erlangen-Nürnberg, BRD both for his valuable discussions and for facilitating access to the CORINA, PETRA, SURFACE, and KMAP programs. The financial support of the KBN Warsaw under grants no: KBN – 4T09A 088 25 3T09A 01127 and PBZ 040 P04/08 is gratefully acknowledged. RG thanks Foundation for Polish Science for an individual grant.

Experimental

The SOM-CoMSA [9], s-CoMSA [24], grid-4D-QSAR [23] and SOM-4D-QSAR [17] procedures were described in the cited previous publications, respectively.

Model builders

All the experimental data, i.e. biological activities for the dihydrofolate reductase inhibitors were extracted from ref. [23] and are given in Table 1

Kohonen mapping

The competitive Kohonen strategy [19] 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).

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

A projection of the molecular electrostatic potential (MEP) value 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 the 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 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 molecular surfaces. The superimposition was performed as shown in Table 2. In practice, 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 from *10x10* to *30*x*30* 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 10^2, 20^2 and 30^2 element vectors.

4D QSAR calculation

We used Hopfinger's spatial grid system [23] for coding molecules. The molecules after AM1 (Austin model 1) optimization were used as initial structures in the molecular dynamic simulation (MDs). Each 3D structure is the starting point in generating conformational ensemble profile (CEP). Molecular dynamics was performed using the Sybyl software [28] with standard Tripos force field. 2500 conformations were sampled for each analogue. Partial atomic charges were calculated using the semiempirical AM1 Hamiltonian (HYPERCHEM package [29]). The alignment of the molecules was the next step of the 4D-QSAR analysis. We aligned the molecules according to the previous rules of the Hopfingers' study [23]. Individual conformers are placed in the grid cell space surrounding the aligned compounds. We applied cubic grid lattice of 20 Å on each side with grid cell resolution of 1, 2 or 0.5 Å, respectively. Different types of grid cell occupancy descriptors (GCODs) were considered and calculated for the indicated atoms referred to as interaction pharmacophore elements (IPE). Apart from, the GCODs used by Hopfinger et al. [23], we applied in our current work the absolute charge occupancy (A_q) for the chosen IPE atoms of compound c defined as

$$A_q(c,i,j,k,N) = \sum_{t=0}^{T} O_t(c,i,j,k) \times q/m \tag{2}$$

where m means the number of the atoms of compounds, c present in the cell (i,j,k) at time t, q means the sum of partial atoms of charges present in some cell at time t, T is the length of the time in MDs. N is the number of sampling MDs steps. The joint (J_q) and self charge occupacy (S_q) with the most active reference compound R were defined after following equations:

$$J_q(c,i,j,k,N) = \sum_{t=0}^{T} O_t(c,i,j,k) \cap O_t(R_q,i,j,k) \times q/m \tag{3}$$

$$S_q(c,R,i,j,k,N) = \sum_{t=0}^{T} \{O_t(c,i,j,k) - [\sum_{t=0}^{T} O_t(c,i,j,k) \cap O_t(R,i,j,k)]\} \times q/m \tag{4}$$

We used the MATLAB [30] environment to program the calculation of the above mentioned descriptors. The Partial Least Squares (PLS) method with variable elimination was used to estimate the relationship between independent variables (GCODs) and corticosteroid binding globulin (CBG) affinity.

Calculation of the molecular surface (s-COMSA) descriptors based on virtual cubic grid

For the calculation of shape descriptors we applied formalism similar to Hopfinger's 4D-QSAR grid coding system using the absolute type descriptors, as given by the above mentioned equations. However, unlike in 4D QSAR our method compares single conformers. Thus, each 3D molecular representation is placed in its own virtual cubic grid and molecular surface is calculated, respectively. The electrostatic potential is calculated for the points randomly sampled on the molecular surface and a mean value of the electrostatic potential corresponding to the respective points found in each grid cell is used to describe this cell. Grid cells are unfolded into vectors and vectors describing all molecules of the series are aligned into a matrix. Grid cells that are empty for all molecules in the series analyzed are eliminated and the resulted matrix was used for further calculations using the PLS method.

PLS analysis

Obtained vectors were processed by the PLS analysis with a leave-one-out cross-validation procedure. The PLS procedures were programmed within the MATLAB environment (MATLAB) [30].

A PLS model was constructed for the centered data and its complexity was estimated on the basis of 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 on the basis of the model with one, two, three etc. factors. The Root Mean Square Error of CV for the model with j factors is defined as:

$$\mathrm{RMSECV_j} = \sqrt{\frac{\sum_i (\mathrm{obs_i} - \mathrm{pred_{i,j}})^2}{\mathrm{m}}} \tag{5}$$

where obs denotes the assayed value; pred - predicted value of 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 $q_{cv}{}^2$

$$q^2_{CV} = 1 - \frac{\sum (obs_i - pred_i)^2}{\sum (obs_i - mean(obs))^2} \tag{6}$$

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

$$s = \sqrt{\frac{\sum (obs_i - pred_i)^2}{m - k - 1}} \tag{7}$$

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

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

The quality of external predictions was measured by the Standard Deviation of Error of Prediction (SDEP) parameter:

$$SDEP = \sqrt{\frac{\sum_i (pred_i - obs_i)^2}{n}} \tag{8}$$

where *pred* – predicted value, *obs* – observed value.

References and Notes

1. Kolb, H.C.; Finn, M.G.; Sharpless, K.B. Click Chemistry: Diverse Chemical Function from a Few Good Reactions. *Angew. Chem., Int. Ed. Engl.* **2001**, *40*, 2004-2021.
2. Buden, F.R.; Winkler, D.A. Robust QSAR models using Bayesian regularized neural networks. *J. Med. Chem.* **1999**, *42*, 3183-3187.
3. Schneider, G.; Wrede, P. Artificial neural networks for computer-based molecular design. *Prog. Biophys. Mol. Biol.* **1998**, *70*, 175-222
4. Polanski, J. Molecular shape analysis. In: *Handbook of Chemoinformatics*; Gasteiger J. (ed.); Wiley-VCH Verlag: Weinheim, **2003**; pp. 302-319.
5. 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.
6. Lucic, B.; Trinajstic, N. Multivariate Regression Outperforms Several Robust Architectures of Neural Networks in QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 121-132.
7. Lucic, B.; Trinajstic, N.; Sild, S.; Karelson, M.; Katritzky, A. R. A New Efficient Approach for Variable Selection Based on Multiregression: Prediction of Gas Chromatographic Retention Times and Response Factors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 610-621.
8. Lucic, B.; Amic, D.; Trinajstic, N. Nonlinear Multivariate Regression Outperforms Several Concisely Designed Neural Networks in QSPR Modeling. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 403-413
9. Polanski, J.; Walczak, B. The comparative molecular surface analysis (CoMSA): a novel tool for molecular design. *Comp. Chem.* **2000**, *24*, 615- 625.
10. Polanski, J.; Gieleciak, R.; Bak, A. 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. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 184-191.
11. Polanski, J.; Gieleciak, R. The comparative molecular surface analysis (CoMSA) with modified uninformative variable elimination-PLS (UVE-PLS) method: application to the steroids binding the aromatase enzym. *J. Chem. Inf. Comput. Sci.* **2003,** ***43,*** 656-666.
12. Polanski, J.; Gieleciak, R.; Wyszomirski, M. Comparative molecular surface analysis (CoMSA) for modeling dye-fiber affinities of the azo and antraquinone dyes, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1754-1762.
13. Polanski, J.; Gieleciak, R.; Wyszomirski, M. Mapping dye pharmacophores by the comparative molecular surface analysis (CoMSA): application to heterocyclic monoazo dyes. *Dyes Pigm.* **2004**, *62*, 63-78.

14. Polanski, J.; Gasteiger, J.; Jarzembek, K. Self - Organizing neural networks for screening and development of novel artificial sweetener candidates. *Combin. Chem. High Throughput Screen.* **2000**, *3*, 481-495.
15. Polanski, J.; Gieleciak, R. Comparative molecular surface analysis: a novel tool for drug design and molecular diversity studies, *Mol. Diversity* **2003**, *7*, 45-59.
16. Polanski, J. Self-organizing neural networks for pharmacofore mapping. *Adv. Drug Deliv. Rev.* **2003**, *55*, 1149-1162.
17. Polanski, J.; Bak, A. Modeling steric and electronic effects in 3D and 4D-QSAR schemes: Predicting benzoic pK_a values and steroid CBG binding affinities. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2081-2092.
18. Polanski, J.; Gieleciak, R.; Bak, A. Probability issues in molecular design: Predictive and modeling ability in 3D-QSAR schemes. *Comb. Chem. High T. Scr.* In press
19. Kohonen, T. *Self-Organization and Associative Memory*, 3rd Edition, Springer Verlag: Berlin, **1989**.
20. Zupan, J.; Gesteiger, J. *Neural Networks in Chemistry and Drug Design, 2nd Edition*; Wiley–VCH: Weinheim, **1999**.
21. Melssen, W.J.; Smits, J.R.M.; Buydens, L.M.C.; Kateman, G. Tutorial: Using artificial neural networks for solving chemical problems. Part II. Kohonen self-organising feature maps and Hopfield networks, *Chemometer. Intell. Lab. Syst.* **1994**, *23*, 267-291.
22. Kohonen, T. The Self-Organizing Map (SOM), http://www.cis.hut.fi/projects/somtoolb.shtml.
23. Hopfinger, A.J.; Wang, S.; Tokarski, J.S.; Jin, B.; Albuquerque, M.; Madhav, P.J.; Duraiswami, C. Construction of 3D QSAR models using the 4D QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509-10524.
24. Polanski, J.; Gieleciak, R.; Magdziarz, T. The grid formalism for the comparative molecular surface analysis: application to the CoMFA benchmark steroids, azo dyes and HEPT derivatives. *J. Chem. Inf. Comput. Sci.* In press
25. Centner, V.; Massart, D. L.; de Noord, O.E.; de Jong, S.; Vandeginste, B. M. V.; Sterna, C. Elimination of uninformative variables for multivariate calibration. *Anal. Chim. Acta.* **1996**, *330,* 1-17.
26. Pilizota, T.; Lucic, B.; Trinajstic, N. Use of variable selection in modeling the secondary structural content of proteins from their composition of amino acid residues *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 113-121
27. Gasteiger, J. Match3D; KMAP for the information see: http://www2.ccc.uni-erlangen.de.
28. Sybyl 6.5. program, available from the Tripos Inc., St. Louis, MO, USA: http://www.tripos.com.
29. *HyperChem 5.0*, available from HyperCube Inc., Gainesville, FL, USA: http://www.hyper.com.
30. *MATLAB 6.5*, available from The Mathworks Inc., Natick, MA, USA, http://www.mathworks.com.