J Mol Model (2009) 15:41–51
DOI 10.1007/s00894-008-0373-1

ORIGINAL PAPER

Receptor independent and receptor dependent CoMSA modeling with IVE-PLS: application to CBG benchmark steroids and reductase activators

Tomasz Magdziarz · Pawel Mazur · Jaroslaw Polanski

Received: 18 June 2008 / Accepted: 22 September 2008 / Published online: 21 October 2008

Abstract Comparative molecular surface analysis (CoMSA) with robust IVE-PLS variable elimination if tested for the benchmark CBG steroid series provides highly predictive RI 3D QSAR models, but failed however to model the activity of sulforaphane (SP) activators of quinone reductase. The application of the SP poses obtained from multipose molecular docking to model the RD IVE-PLS CoMSA resulted in a predictive form. This model indicated lipophilic potential as the activity determinant. The individual molecular surface areas of the highest contribution to the SP activity was identified and visualized by CoMSA contour plots.

Keywords Comparative molecular surface analysis · CoMSA · IVE-PLS · Receptor dependent 3D QSAR · Reductase activators · Sulforaphane

Introduction

Quantitative structure activity relationship (QSAR) is an approach mapping chemical structure to properties that should convert molecular data to drugs by property prediction and design. A significant development can be observed along the last decades in this method. A traditional Hansch analysis based on the logP and Hammett constant has been supplemented with 3D QSAR methods that can account for 3D structure, conformational dynamics and finally receptor data and solvation effects. However modeling interactions of chemical molecules in biological systems still provides highly noisy data, which makes activity predictions a roulette risk. This can be classified as the data, superimposition, molecular similarity, conformational, and molecular recognition noise [1]. Molecular recognition uncertainty in traditional receptor independent (RI) m-QSAR cannot be removed but by the inclusion of the receptor data. However, modeling ligand-receptor interactions is a complex computational problem, which limited the development of the receptor dependent (RD) m-QSAR. It is just recently that RD m-QSAR methods became popular [2]. The idea started as early as the 90' from the application of the CoMFA – like molecular interaction force filed (MIF) and the GRID method to investigate the binding pockets of the receptors [3]. Further development resulted in the RD 4D, 5D and 6D QSAR methods or membrane interactions (MI) QSAR [4–8].

In the majority of applications 3D QSAR describes the RI model sampled from the single conformation representations. A 3D QSAR query in the Pubmed database provides 742 hits (CoMFA - 772; CoMFA AND 3D QSAR 407). The latter numbers illustrate the predominance of the CoMFA concept in the ligand based multidimensional QSAR [2]. It is however not only an advantage of the method but the availability of the CoMFA software that decides that CoMFA outnumbers other approaches. This *has limited both the evaluation and use of other QSAR methodologies* [8] and a number of other multidimensional descriptors can be used for modeling RI and RD 3D QSARs [2].

Human NAD(P)H quinone oxidoreductase is an enzyme overexpressed in a variety of solid tumors, which makes it an interesting target for anticancer drugs. Quinone oxidoreductase plays a protective antioxidant role being also capable of bioactivation of a variety of prodrugs to their cytotoxic species. Several novel inhibitor series of this enzyme were reported recently [9]. A virtual screening among the more than 700,000 molecule compound library

T. Magdziarz · P. Mazur · J. Polanski (✉)
Department of Organic Chemistry, Institute of Chemistry,
University of Silesia,
PL-40–006 Katowice, Poland
e-mail: polanski@us.edu.pl
URL: http://prac.us.edu.pl/~zchorg

was performed to identify the potential ligand of 1D4A reductase. This docking approach resulted in the design of novel active structures; however, no correlation between the calculated and measured binding energies for the analyzed compounds was observed [10]. Sulforaphanes (SPs) are compounds activating quinone reductase enzyme closely as the second phase of a detoxification. Thus, SPs can be applied as chemopreventive agents and a number of investigations have been reported for these compounds. However, only few studies reported structure activity relationships for the series, which indicates that sulforaphane itself is the most potent inducer [11]. Previous experimental studies failed to indicate the molecular basis for the SP activatory activity [12, 13], which inspired us to investigate this effect *in silico* using molecular docking. Although we failed to correlate the SP activity to the docking scoring functions, the application of the activator-enzyme complex for the simulation of the reductase inhibition indicated an interesting enhancement mechanism in which a formation of the SP- reductase complex modifies binding cavity of the enzyme exposing the TYR 128 residue for a further substrate binding [14], which is a rare example of the manipulation of the drug-enzyme complex for a simulation of a further enzyme behavior. Similar modeling study has been described for HIV-1 integrase [15].

We have described previously the comparative molecular surface analysis (CoMSA) [16–24] which was then supported by the robust variable elimination method [25]. This was however used only in the traditional RI mode. In the present paper we attempt to extend the application of CoMSA method with iterative variable elimination (IVE-PLS) to the RD modeling of the activatory activity of the series of SP compounds interacting with quinone oxidoreductase. Since, we modified here the original IVE-PLS method [16, 20] we also tested the performance of the method during the application to the benchmark CBG steroid series.

Data sets and methodology

Data sets

All compounds examined in the present study were reported previously in the literature. The CBG steroid benchmark series data, molecules **s1-s31**, were reported according to reference [17]. The SP data, molecules **r1-r10**, were extracted form refs. [11–13, 26]. The data is presented in Tables 1 and 2, respectively.

Molecular modeling and docking

Molecular modeling was conducted using the Sybyl/Tripos or CCG MOE software packages running on an Intel Pentium based machine with the GNU/Linux CentOS operating system. The initial geometry of CBG steroids was optimized using standard Tripos force field (POWELL method) with 0.005 kcal/mol energy gradient convergence criterion and a distant dependent dielectric constant. Partial atomic charges were calculated using the Gasteiger-Marsili method implemented in Sybyl. The set was superimposed by MATCH 3D program and as a superimposition template compounds **s6** were used. SPs were modeled using the MOE software. The initial geometry was optimized using the MMFF94x force field with 0.01 kcal/mol gradient convergence criterion and the force field partial charges were calculated.

Alternatively, compounds **r1-r10** were modeled within the receptor structure 1D4A PDB [27] using the MOE docking protocol with the Alpha Triangle placement option and the London dG scoring function. Missing hydrogen atoms were added to the receptor structure and a titration to the protonation state at pH 7.4 was performed. The potential docking sites were identified using the Site Finder procedure and the four mostly populated sites were used for further docking. For each ligand 100 poses were saved yielding overall 1000 poses, out of which 956 poses were placed in the first potential site. Poses yielding the highest score were chosen for further CoMSA analyses - one pose per ligand. Molecule **r4** for which the first three poses of the highest scoring have been docked apart from the bundle, was modeled in QSAR in the forth pose.

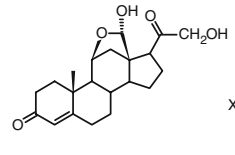
Comparative molecular surface analysis

Molecular shape descriptors in present work were calculated by grid formalism of the s-CoMSA method. Thus, each 3D molecular representation is placed in its own virtual cubic grid and molecular surface is calculated, respectively. The electrostatic (*ep*) and/or the lipophilic (*lipo*) potentials are calculated for the points randomly sampled on the molecular surface and a mean value of the potential corresponding to the respective points found in each grid cell (or other value) is used to describe this cell. Calculated values are unfolded into vectors and vectors describing all molecules of the series are aligned in to a matrix. Columns corresponding to grid cells that are empty for all molecules in the series are eliminated. The resulting matrix is used for further calculations using the PLS and IVE-PLS methods.

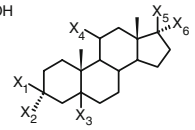
Iterative variable elimination IVE-PLS method

IVE-PLS method is an iterative extension of the uninformative variable elimination (UVE-PLS) algorithm originally proposed by Centner et al. [28] as a possible improvement of the PLS procedure. The main idea of UVE-PLS is to reduce the number of the redundant

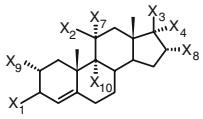
Table 1 Steroid structures and the CBG affinity data [20]



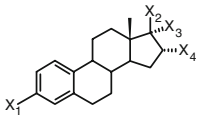
SA



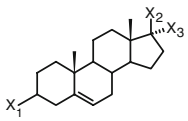
SB



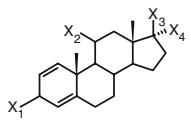
SC



SD



SE



SF

Nr	S	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	CBG
s1	SA											-6.279
s2	SB	OH	H	H[a]	H	OH	H					-5.000
s3	SE	OH	OH	H								-5.000
s4	SC	=O	H	=O				H	H	H	H	-5.763
s5	SB	H	OH	H[a]	H	=O						-5.613
s6	SC	=O	OH	$COCH_2OH$	H			H	H	H	H	-7.881
s7	SC	=O	OH	$COCH_2OH$	OH			H	H	H	H	-7.881
s8	SC	=O	=O	$COCH_2OH$	OH				H	H	H	-6.892
s9	SE	OH	=O									-5.000
s10	SC	=O	H	$COCH_2OH$	H			H	H	H	H	-7.653
s11	SC	=O	H	$COCH_2OH$	OH			H	H	H	H	-7.881
s12	SB	=O		H[a]	H	OH	H					-5.919
s13	SD	OH	OH	H	H							-5.000
s14	SD	OH	OH	H	OH							-5.000
s15	SD	OH	=O		H							-5.000
s16	SB	H	OH	H[b]	H	=O						-5.255
s17	SE	OH	COMe	H								-5.255
s18	SE	OH	COMe	OH								-5.000
s19	SC	=O	H	COMe	H			H	H	H	H	-7.380
s20	SC	=O	H	COMe	OH			H	H	H	H	-7.740
s21	SC	=O	H	OH	H			H	H	H	H	-6.724
s22	SF	=O	OH	$COCH_2OH$	OH							-7.512
s23	SC	=O	OH	$COCH_2OCOMe$	OH			H	H	H	H	-7.553
s24	SC	=O	=O	COMe	H				H	H	H	-6.779
s25	SC	=O	H	$COCH_2OH$	H			OH	H	H	H	-7.200
s26	SC[c]	=O	H	OH	H			H	H	H	H	-6.144
s27	SC	=O	H	COMe	OH			H	OH	H	H	-6.247
s28	SC	=O	H	COMe	H			H	Me	H	H	-7.120
s29	SC[c]	=O	H	COMe	H			H	H	H	H	-6.817
s30	SC	=O	OH	$COCH_2OH$	OH			H	H	Me	H	-7.688
s31	SC	=O	OH	$COCH_2OH$	OH			H	H	Me	F	-5.797

[a] 5-α
[b] 5-β
[c] H instead Me at the C_{10}

Table 2 Sulforaphanes structures and reductase activation rate [11–13, 26]

Nr/ name	Structure	Activation rate A [μM/l]	Activation rate pA (-log A)
r1	$CH_3(CH_2)_5NCS$	15 [11]	-1.1761
r2	$CH_3(S{=}O)(CH_2)_4NCS$	0.2 [11]	0.6989
r3	$CH_3(C{=}O)(CH_2)NCS$	0.2 [11]	0.6989
r4	$CH_3(CH_2)_3(C{=}O)(CH_2)_4NCS$	2.0 [11]	-0.3010
r5	$CH_3(S{=}O)(CH_2)_3NCS$	0.4 [11, 26]	0.3971
r6	$CH_3S(C{=}O)(CH_2)_4NCS$	2.8 [26]	-0.4472
r7	$CH_3O(C{=}O)(CH_2)_4NCS$	2.8 [26]	-0.4472
r8	$N{\equiv}C(CH_2)_4NCS$	2.0 [26]	-0.3010
r9	$CH_3(S{=}O)(CH_2)_5NCS$	1.6 [12, 13]	-0.2041
r10	$CH_3(C{=}O)(CH_2)_4NCS$	0.5 [12, 13]	0.3010

variables included in the final model. The UVE algorithm based on the analysis of the regression coefficients calculated by the PLS method. The PLS method allows presenting the relation between the **Y** answer and the **X** predictors in a form of

$$\mathbf{Y} = \mathbf{X}\boldsymbol{b} * \boldsymbol{e} \quad (1)$$

where $\boldsymbol{b}$ is a vector of the regression coefficients and $\boldsymbol{e}$ is the vector of the errors. Thus, the UVE algorithm analyzes a value of $\boldsymbol{t}$ called stability that is calculated on the basis of the $\boldsymbol{b}$ coefficients of the PLS Eq. 1. The $\boldsymbol{t}$ score for the variables is given by Eq 2:

$$\boldsymbol{t} = \mathrm{mean}(\mathbf{B})/\mathrm{std}(\mathbf{B}) \quad (2)$$

where **B** is a matrix of $\boldsymbol{b}$ coefficients obtained during the leave-one-out cross-validation procedure and mean and std are mean and standard deviation values, respectively.

Then, only the variables of the "relative" high $\boldsymbol{t}$-value are included in the final PLS model. In order to estimate the cutoff level, the artificial random number noise is created (the level of the noise is 10^{-10} of the original variable order) and added as additional columns into the matrix of the original variables.

We have modified this procedure replacing a single step procedure with the iterative algorithm, which is based on the absolute value abs(mean(**B**)/std(**B**)) as a criterion to identify variables to be eliminated. To distinguish this procedure, we named this method as the iterative variable elimination (IVE-PLS). This procedure includes the following steps:

1. Standard PLS analysis applied to analyze the matrices yielded from the s-CoMSA procedure with the leave-one-out cross-validation to estimate the performance of the PLS model (q^2),
2. Elimination of the matrix column of the lowest abs(mean(**B**)/std(**B**)) value,
3. Standard PLS analysis of the new matrix without the column eliminated in step 2,
4. Iterative repetition of the steps 1–3 to maximize the LOO CV q^2 parameter.

The detailed procedure for several IVE-PLS versions was described in ref. [25] where several robust measures of the mean operator in criterion 2 were tested. In the current version we applied the robust IVE version which defines the stability criterion by equation:

$$\boldsymbol{t} = \mathrm{median}(\mathbf{B})/\mathrm{iqr}(\mathbf{B}) \quad (3)$$

where median and iqr are median value and interquartile range respectively.

Unlike in standard PLS, in this method a number of PLS components are usually truncated at an arbitrarily decided level A_{max} that was always lower or equal to an optimal number of latent PLS variables. Our experience indicates that such a truncation allows one to obtain highly predictive models. The detailed study on the influence of the number of the truncation extent on the model quality can be found

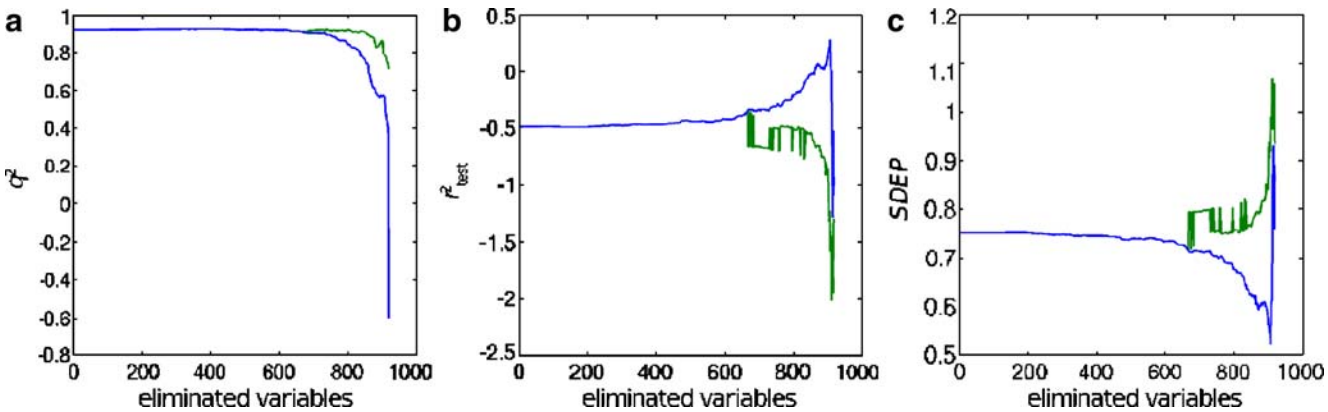


Fig. 1 The variable elimination IVE-PLS profile in the s-CoMSA modeling of the CBG steroid series by maximization of q^2_{cv} for the training set **s1-s21** (**a**) accompanied by the r^2_{test} (**b**) and SDEP (**c**) profiles for the test set **s22-s31**, details in text

Table 3 The performances of RI CoMSA modeling of the CBG steroid series

Entry	Training /Test set	A_{max}	q^2_{cv}	*SDEP*	r^2_{test}	IVE-PLS			Number of variables Initial/Final
						q^2_{cv}	*SDEP*	r^2_{test}	
1	s1-s21/s22-s31	1	0.92	0.75	-0.49	0.93	0.75	-0.47	919/593
2	s1-s12 s23–31 /s13-s22	1	0.68	0.47	0.83	0.71	0.41	0.87	919/245

in reference [25]. The performance of the IVE-PLS method without component truncation has been recently compared by Grohmann to the performances of other robust PLS methods [29].

We used the standard cross-validated PLS performances, namely, q^2_{cv}, r^2_{test} and *SDEP* to measure the quality of the PLS models [25]. Moreover, in the so called Y-randomization procedure we further validated model quality. Thus, the activity (**Y** answer) was randomly permuted in a series of experiments and the whole IVE-PLS was repeated to compare the resulted q^2_{cv} values of the pseudomodels with this of the real model. In particular, we simulated here 1000 Y-randomized pseudomodels.

Drug design toolbox

The UVE and IVE procedures were programmed within the MATLAB environment (MATLAB) and were included in the drug design toolbox (DDT) developed in our group [30]. DDT consists of two software layers. The first layer performs all calculations and basic input – output operations including importing and exporting molecular data. All first layer functions can be accessed by MATLAB command line and can be easily linked to other MATLAB functions and scripts. The second layer is a graphical user interface. All calculations run by the second layer are accomplished by appropriate first layer functions which can be used as a stand alone command line toolbox. The software is capable of importing and exporting molecular data from/to mol2 Sybyl [31] and ctx CACTVS [32] files. However, during calculations DDT operates its own molecular format IQF (internal QSAR format). Similarly, data resulted form QSAR modeling are stored in DDT format, namely UQS (universal QSAR structure). Both formats, IQF and UQS, are XML based and can be saved as plain text files or in MATLAB binary format. Moreover, in order to organize and simplify batch operations on huge molecular data DDT can create special directories QDB (QSAR data base) containing molecular series in IQF formats. Such directories can be accessed by DDT batch routines speeding up operations on huge molecular series.

The toolbox allows a generation of the van der Waals molecular surfaces and a calculation of the electrostatic potential. However, partial charges have to be calculated using a third party software. There is the ALOGP [33] method implemented and the lipophilic potential can also be calculated using the Audry method [34].

Molecular descriptors can be calculated by grid (s-CoMSA) and SOM (SOM-CoMSA) versions, though, the freeware Kohonen SOM toolbox [35] is required to use the latter method. DDT makes available several data preprocessing protocols. Quantitative modeling can be realized by PCR and PLS methods. Both UVE-PLS and the several versions of IVE-PLS were implemented in DDT. A variety of coloring maps are available for molecular visualization and displaying CoMSA contour plots, as described in previous publications [1, 16, 36, 37]. DDT can be downloaded as a freeware from our internet site [30].

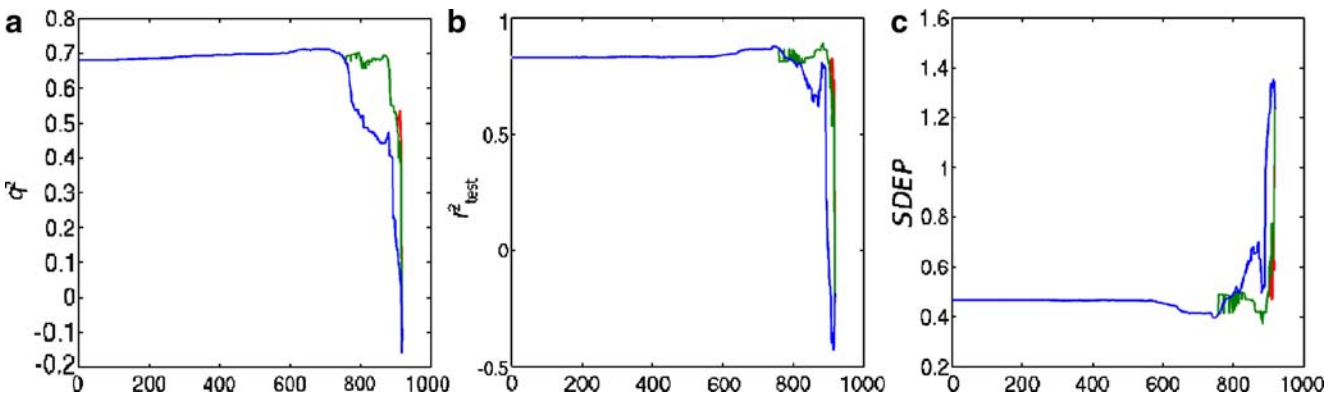


Fig. 2 The variable elimination IVE-PLS profile in the s-CoMSA modeling of the CBG steroid series by maximization of q^2_{cv} for the training set **s1-s12** and **s23-s31** (**a**) accompanied by the r^2_{test} (**b**) and SDEP (**c**) profiles for the test set **s13-s22**, details in text

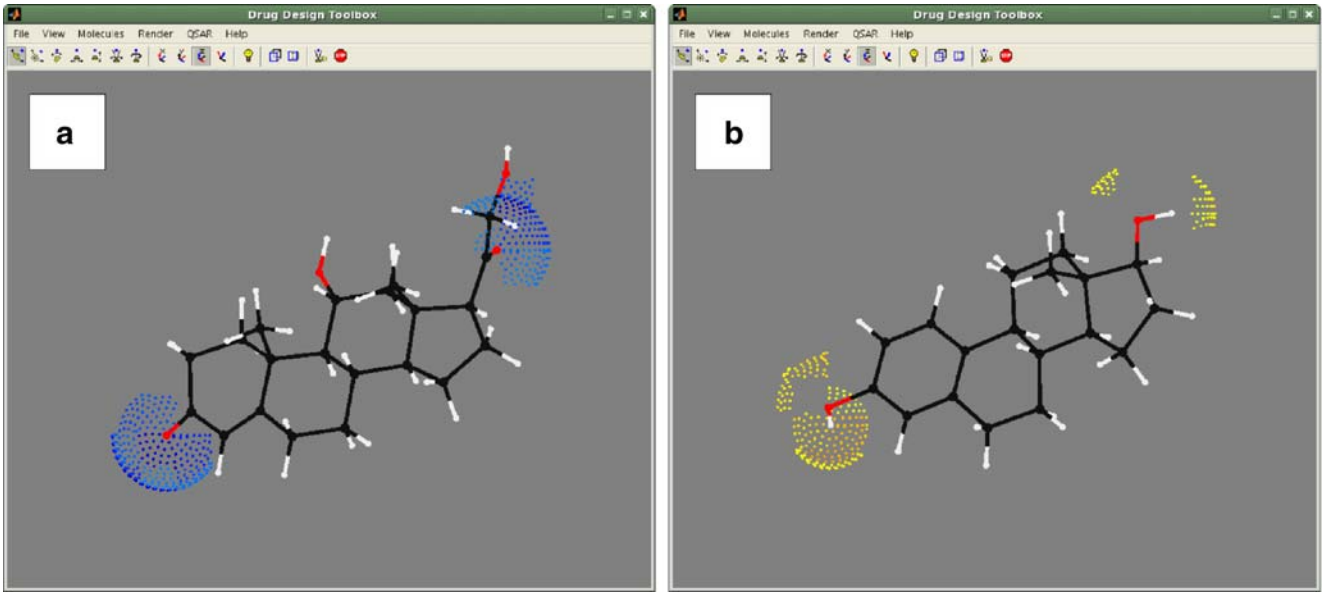


Fig. 3 The s-CoMSA contour plots for the CBG steroids **s6** – high affinity (**a**) and **s13** – low affinity (**b**). Colors code a contribution of molecular surface *ep* potential into a final IVE-PLS model. Blue increases while red and yellow decreases the activity value. For more clear illustration the additional data filter is applied to eliminate variables of the lowest contribution, i.e., only 50% of the highest contribution variables surviving IVE-PLS are shown

Result and discussion

RI s-CoMSA for the steroid benchmark series

The original data of the series of steroids complexing corticosteroid binding globulin (CBG) come from publications by Mickelson et al. [38], Westphal [39], and Dunn et al. [40]. Due to the rigid steroid skeleton, this series is used in molecular design as a benchmark measuring the performance of new methods. However, a number of early publications analyzing these series include several errors within the molecular structures [17, 24]. This was corrected by Wagener at al. [41].

Table 4 The performances of RI and RD CoMSA modeling of the reductase activation rate by SP compounds

Entry	Model	A_{max}	q^2_{cv}	q^2_{cv} [a]	Number of variables Initial/Final
1	RI s-CoMSA ep	3	-1.15	-0.72	671/316
2	RI s-CoMSA lipo	3	0.16	0.49	671/180
3	RD s-CoMSA ep	1	-0.73	-0.15	1074/188
4	RD s-CoMSA ep	2	-0.73	-0.01	1074/39
5	RD s-CoMSA ep	3	-0.73	0.01	1074/41
6	RD s-CoMSA lipo	1	0.33	0.66	1074/326
7	RD s-CoMSA lipo	2	0.33	0.67	1074/538
8	RD s-CoMSA lipo	3	0.33	0.77	1074/404

[a] with IVE-PLS

As reported in previous publications we distributed the CBG steroids into the training **s1**-**s21** and test sets **s22**-**s31**. In Fig. 1 we presented the q^2_{cv} profile during IVE-PLS s-CoMSA modeling with a number of PLS components

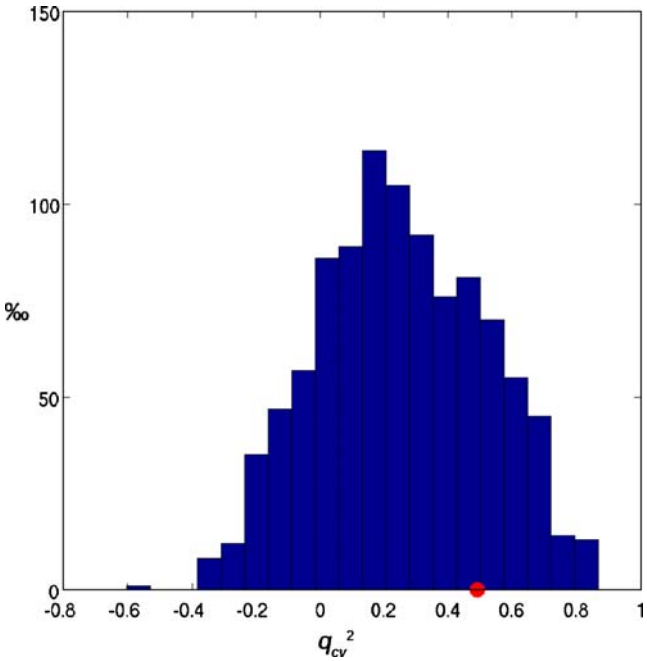


Fig. 4 The Y-randomization pseudomodels of the IVE-PLS RI s-CoMSA of the SP series: Table 4 entry 2. The red dot indicates the q^2_{cv} values for the correct activity model

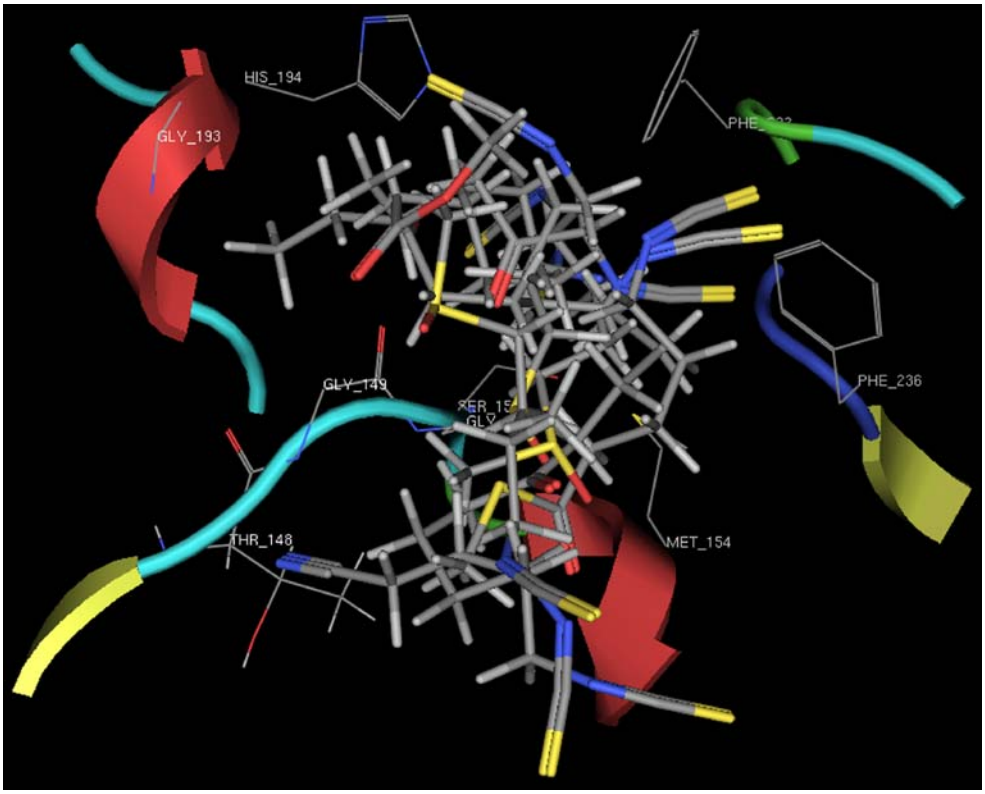


Fig. 5 SP compounds in the molecular superimposition poses resulted by docking simulation from the reductase 1DA4

A_{max} truncated at 1 or 2, respectively. The method allowed us to obtain a highly predictive model with q^2_{cv} amounting to 0.93 (for molecules **s1-s21**) for A_{max}=1 with 593 out of 919 (ca. 65%) variables surviving the IVE-PLS data elimination – see Table 3. This compares advantageously, for example, to the Quasar model with q^2_{cv} amounting to 0.90 [42]. However, the predictive ability for the test set molecules **s22-s31** r^2_{test} amounts to -0.47. This indicates that the model does not reach the predictive values for the test set. The relationship plotted in Fig. 1b can be transformed into the standard deviation of external predictions (SDEP) error, which is shown in Fig. 1c. This reveals that the initial SDEP value amounts to 0.75, which falls within the range of the values described in the majority of publications, where SDEP ranges from 0.7 to 0.8 [16, 17]. In particular this significantly outperforms the CoMFA SDEP taking a value of 0.837 [43]. Moreover, the robust CoMSA architecture allowed the SDEP to decrease after IVE variable elimination, i.e., at its minimal level to a value of ca. 0.5. This decrease is however accompanied by the decrease in the q^2_{cv} rate to a value of ca. 0.6.

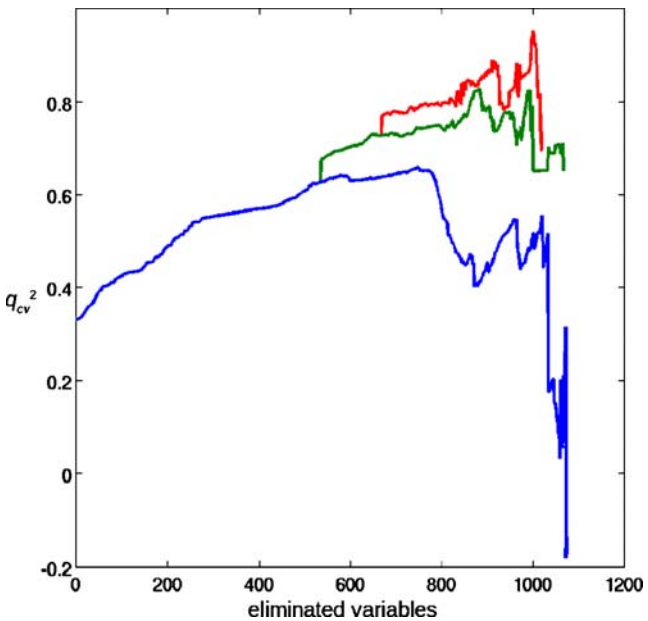


Fig. 6 The variable elimination IVE-PLS profile in the s-CoMSA modeling of the SP series. Colors code a value of A_{max}: blue 1; green 2 and red 3, respectively

The analysis discussed above illustrates a fact that the distribution of the CBG within training and test sets that can usually be found in the literature, is non-representative for the analyzed structures and provides non-predictive models for the test set compounds, which was first realized by Kubinyi. Therefore, he recommended another training/test set distribution, namely, test set: **s1-s12** and **s23-s31**/training set: **s13-s22** [17]. After such a correction we obtained the IVE-PLS CoMSA model described by q^2_{cv} amounting to 0.71 (A_{max}= 1) for molecules **s1-s12** and **s23-s31**, and $r^2_{test} = 0.87$ or

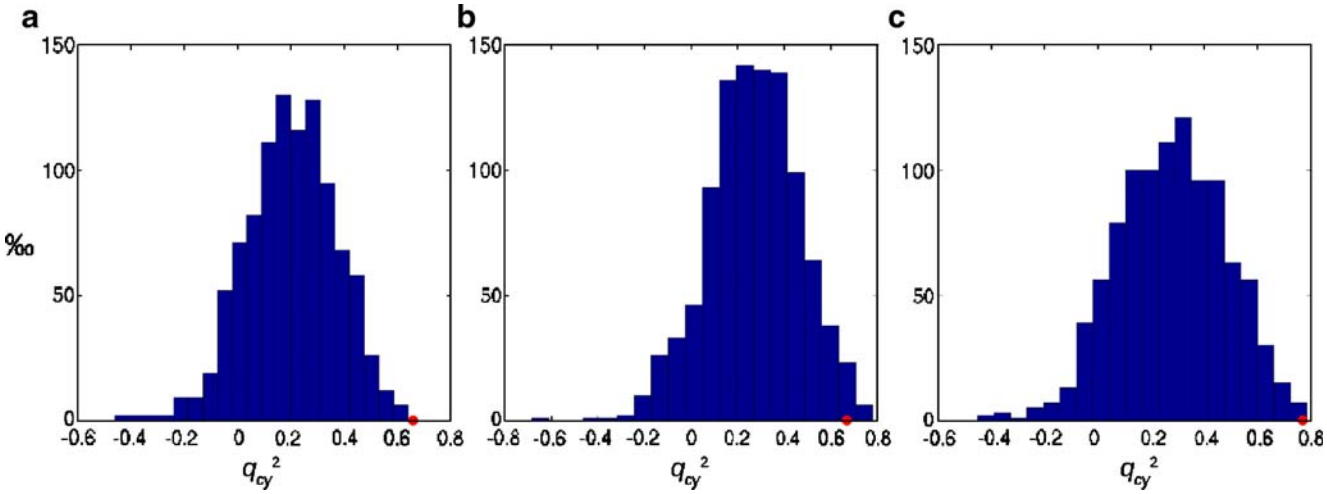


Fig. 7 The Y-randomization pseudomodels of the IVE-PLS s-CoMSA of the SP series: Table 4 entries 6 (**a**), 7 (**b**) and 8 (**c**), respectively. The red dots indicate the q^2_{cv} values for the correct activity models

SDEP=0.41 for molecules **s13-s22** with 27% variables surviving the IVE-PLS elimination – see Table 3. The detailed IVE-PLS CoMSA profiles are shown in Fig. 2. In particular the q^2_{cv} of 0.71 significantly outperforms this of the CoMFA q^2_{cv} that amounts to 0.454 [17].

However, to compare the models with those previously reported in the literature we plotted in Fig. 3 these variables that survive IVE-PLS with standard training/test steroid **s1-s21**/**s22-s31** distribution. This indicates the surface areas deciding the activity of the CBG series. Generally, the molecular surface sectors complies with those reported to be important in the previous publications [16]. This indicates the A and D steroid rings and substitutions as those determining the activity – see Fig. 3.

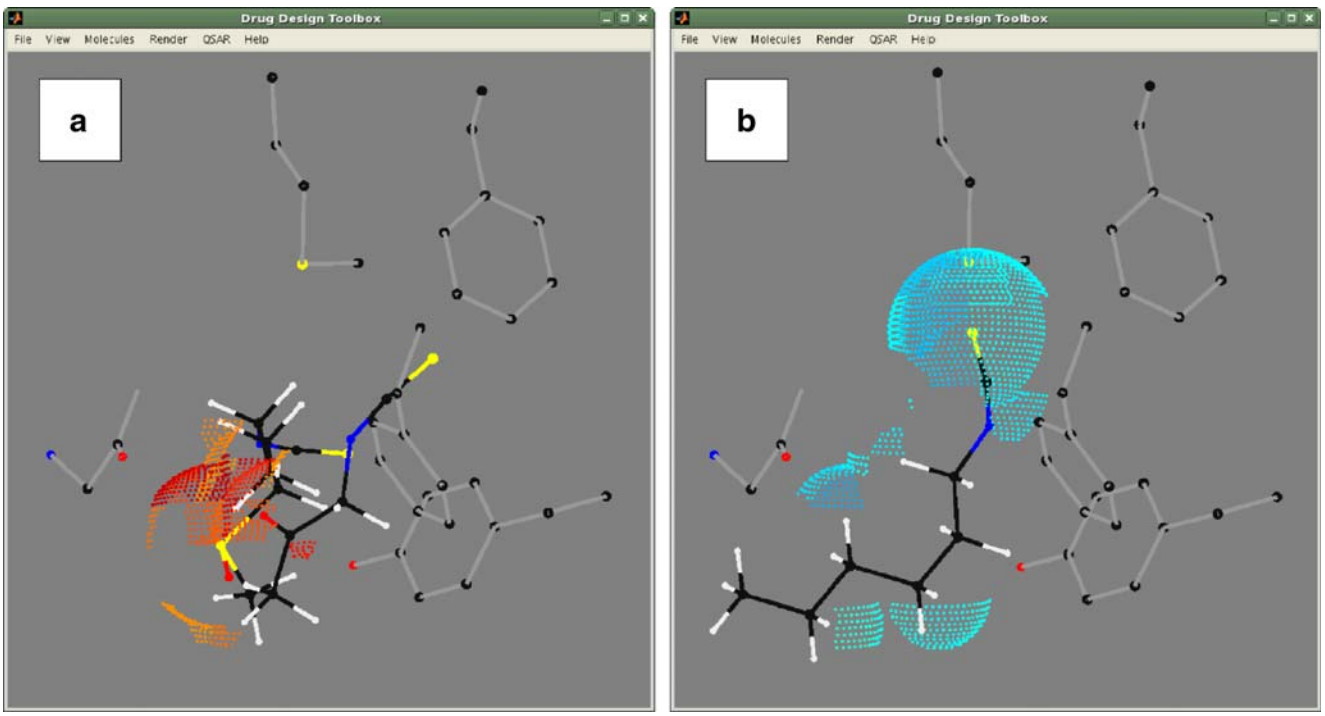


Fig. 8 The RD s-CoMSA contour plots for the SP series **r2** and **r3** – high affinity (**a**), **r1** – low affinity (**b**). The reductase residues Tyr 128, Gly 149, Gly 150, Met 154, Phe 232, Phe 236 are shown in gray. Colors code a contribution of molecular surface lipo potential into a final IVE-PLS model. Blue decreases while red increases the activity value. For more clear illustration the additional data filter is applied to eliminate variables of the lowest contribution, i.e., only 50% of the variables having the highest contribution are displayed

RI and RD CoMSA for chemopreventive sulforaphanes

Irrespective of the tested receptor independent superimposition modes, our efforts to model the RI s-CoMSA failed, as shown in Table 4 – entries 1 and 2. Only non-predictive models can be obtained and a value of 0.16 was the maximal q^2_{cv} performance value which, however, improves in IVE-PLS to a value of 0.49. A value of 0.49 is not high enough to consider the model predictable. Since only ten molecules were available in this study we performed Y-randomization test to validate model quality, i.e., the **Y** answers were permutated to take the random values [44]. This indicates a large chance of model overfitting, as shown in Fig. 4. Thus, the q^2_{cv} parameter calculated for the model with correct activity takes a value of 0.49 which is located in the middle of the q^2_{cv} range of the Y-randomized pseudomodels that oscillates from -0.60 to 0.87. Further improvement of the model cannot be achieved. Thus, we separated the SP compounds from the reductase receptor data in the molecular superimposition poses determined by docking simulation, as shown in Fig. 5 and use this for modeling the RD s-COMSA. The results are shown in Table 4, entries 3 to 8 and Fig. 6, 7, and 8. In Fig. 6 we illustrated the IVE-PLS s-CoMSA profiles with different A_{max} levels ranging from 1 to 3. This shows a steady increase of the q^2_{cv} value up to ca. 800 – 1000 eliminated variables depending upon the A_{max} value. In Fig. 7 we presented histograms illustrating the results of the Y-randomization tests. This indicated the predictive ability of model 6 from Table 4 based on *lipo* for which q^2_{cv} performance was higher than any of the q^2_{cv} parameters calculated for the Y-randomized pseudomodels. Vice versa, randomization does not change a low predictive ability of the *ep* models (data not shown here). Thus, our study indicated that lipohilic potential determines binding affinity of the SPs to the quinone reductase.

It is worth mentioning, that IVE-PLS procedure truncated to a single component (A_{max}=1) allowed us to improve the initial model from $q^2_{cv} = 0.33$ to a final value of 0.66. Figure 6 illustrates a profile of q^2_{cv} during the IVE-PLS data elimination process. Although, the higher number of the A_{max} allowed for the larger increase of the q^2_{cv} performances, the randomization indicates higher chances of model overfitting, as compared in Fig. 7, which is an important hint for the IVE-PLS A_{max} protocol.

In Fig. 8 we illustrated the CoMSA contour plots which reveal the areas determining the high and low SP activatory activity. Thus, high affinity **r2** and **r3** compounds (Fig. 8a) are compared with the low activity molecule **r1** (Fig. 8b). Red surface areas increases the activity while blue tends to decrease it. The most important determinants of activity appear in the proximity of Gly 149, Gly 150, and Tyr 128 (left bottom part of Fig. 8a and b) distinguishing between active and inactive compounds. Thus, hydrophobic interactions for high activity molecules come into sight in these locations (Fig. 8b). Vice versa, low activity analogues indicates a completely different *lipo* profile in these areas, as shown for compound **r1** (Fig. 8a). The contour plots in the proximity of Met 154 are less specific although high lipophilic NCS functionality is not favorable in that area. Instead the NCS group located near Phe 236 or Phe 232 seems to be advantageous for the high SP activity.

Conclusions

We described the comparative molecular surface analysis with robust IVE-PLS variable elimination. This method is tested for the benchmark CBG steroid series and provides highly predictive RI models. The same method applied for a series of SP activators of quinone reductase provided nonpredictive RI models. However, the application of the SP poses obtained from multipose molecular docking to model the RD CoMSA IVE-PLS resulted in a predictive form. Moreover, this indicated lipophilic potential as the activity determinant. The individual molecular surface areas of the highest contribution to the activatory activity was identified and visualized by CoMSA contour plots which reveal the areas determining compounds' affinity. The important hints can be concluded from the q^2_{cv} profiles during variable elimination in IVE-PLS. Both for the CBG steroid series and SP compounds the predictive ability of the models depends upon the PLS latent variable truncation level A_{max}. Thus, the lower this value is, the higher the predictive ability of the model in the test set or a better Y-randomization ratio will be observed.

Acknowledgments The authors thank Professor Johann Gasteiger of the University of Erlangen-Nuremberg, BRD for facilitating access to the MATCH 3D program. The financial support of the KBN Warsaw under grant no. R0504303 is gratefully acknowledged.

References

1. Polanski J, Bak A, Gieleciak R, Magdziarz T (2006) Modeling Robust QSAR. J Chem Inf Model 46:2310–2318. doi:10.1021/ci050314b
2. Polanski J Receptor Dependent Multidimensional QSAR for Modeling Drug – Receptor Interactions. Curr Med Chem sent for publication
3. Head RD, Smythe ML, Oprea TI, Waller CL, Green SM, Marshall GR (1996) VALIDATE: A new method for the receptor-based prediction of binding affinities of novel ligands. J Am Chem Soc 118:3959–3969. doi:10.1021/ja9539002
4. Hopfinger AJ, Wang S, Tokarski JS, Jin B, Albuquerque M, Madhav PJ et al (1997) Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. J Am Chem Soc 119:10509–10524. doi:10.1021/ja9718937

5. Vedani A, Briem H, Dobler M, Dollinger H, McMasters DR (2000) Multiple-conformation and protonation-state representation in 4D-QSAR: the neurokinin-1 receptor system. J Med Chem 43:4416–4427. doi:10.1021/jm000986n
6. Vedani A, Dobler M, Lill MA (2005) Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. J Med Chem 48:3700–3703. doi:10.1021/jm050185q
7. Lill MA, Vedani A (2006) Combining 4D pharmacophore generation and multidimensional QSAR: modeling ligand binding to the bradykinin B2 receptor. J Chem Inf Model 46:2135–2145. doi:10.1021/ci6001944
8. Esposito EX, Hopfinger AJ, Madura JD (2003) 3D- and nD-QSAR methods. In: Gasteiger J (ed) Handbook of chemo-informatics: from data to knowledge, vol. 4. Wiley-VCH, Weinheim, pp 1576–1599
9. Colucci MA, Moody CJ, Couch GD (2008) Natural and synthetic quinones and their reduction by the quinone reductase enzyme NQO1: from synthetic organic chemistry to compounds with anticancer potential. Org Biomol Chem 6:637–656. doi:10.1039/b715270a
10. Nolan KA, Timson DJ, Stratford IJ, Bryce RA (2006) In silico identification and biochemical characterization of novel inhibitors of NQO1. Bioorg Med Chem Lett 16:6246–6254. doi:10.1016/j.bmcl.2006.09.015
11. Posner GH, Cheon-Gyu C, Green JV, Zhang Y, Talalay P (1994) Design and synthesis of bifunctional isotiocyanate analogs of sulforaphane. Corelation between structure and potency as inducers of anticarcinogenic detoxication enzymes. J Med Chem 37:170–176. doi:10.1021/jm00027a021
12. Misiewicz I, Skupińska K, Kowalska E, Lubiński J, Kasprzycka-Guttman T (2004) Sulforaphane mediated induction of a phase 2 detoxifying enzyme NAD(P)H quinone reductase and apoptosis in human lymphoblastoid cells. Acta Biochim Pol 51:711–721
13. Misiewicz I, Skupińska K, Kasprzycka-Guttman T (2007) Differential response of human healthy lymphoblastoid and CCRF-SB leukemia cells to sulphoraphane and its two analogues: 2-oxohexyl isothiocyanate and alyssin. Pharmacol Rep 59:80–87
14. Mazur P, Magdziarz T, Chilmonczyk Z, Kasprzycka-Guttman T, Misiewicz I, Skupinska J, Polanski J (in press) Receptor Dependent 3D QSAR model of the chemopreventive sulforaphanes activating oxidoreductase. Bioorg Med Chem Lett
15. Savarino A (2007) In-Silico docking of HIV-1 integrase inhibitors reveals a novel drug type acting on an enzyme/DNA reaction intermediate. Retrovirology 4:21. doi:10.1186/1742–4690–4–21
16. Polanski J, Gieleciak R, Magdziarz T (2004) 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 44:1423–1435. doi:10.1021/ci049960l
17. Coats E (1998) The CoMFA steroids as a benchmark dataset for development of 3D QSAR methods. Perspect Drug Discov Des 12/13(14):199–213. doi:10.1023/A:1017050508855
18. Anzali S, Gasteiger J, Holzgrabe U, Polanski J, Teckentrup A, Wagener M (1998) The use of self-organizing neural networks in drug design. Perspect Drug Discov Des 9/10(11):273–299. doi:10.1023/A:1027276425268
19. Polanski J, Gieleciak R, Bak A (2002) The comparative molecular surface analysis (CoMSA) - a nongrid 3D QSAR method by a coupled neural network and PLS system: Predicting pKa values of benzoic and alkanoic acids. J Chem Inf Comput Sci 42:184–191. doi:10.1021/ci010031t
20. Polanski J, Gieleciak R (2003) 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 43:656–666. doi:10.1021/ci020038q
21. Polanski J, Gieleciak R, Wyszomirski M (2003) Comparative molecular surface analysis (CoMSA) for modeling dye-fiber affinities of the azo and antraquinone dyes. J Chem Inf Comput Sci 43:1754–1762. doi:10.1021/ci0340761
22. Polanski J, Gieleciak R (2003) Comparative molecular surface analysis: a novel tool for drug design and molecular diversity studies. Mol Divers 7:45–59. doi:10.1023/B:MODI.0000006536.02970.f0
23. Polanski J, Gieleciak R, Bak A (2004) Probability issues in molecular design: predictive and modeling ability in 3D-QSAR schemes. Comb Chem High Throughput Screen 7:793–807. doi:10.2174/1386207043328292
24. Polanski J, Walczak B (2000) The comparative molecular surface analysis (CoMSA): a novel tool for molecular design. Comput Chem 24:615–625. doi:10.1016/S0097–8485(00)00064–4
25. Gieleciak R, Polanski J (2007) Modeling robust QSAR. 2. Iterative variable elimination schemes for CoMSA: application for modeling benzoic Acid pKa values. J Chem Inf Model 47:547–556. doi:10.1021/ci600295z
26. Zhang Y, Kensler TW, Posner GH, Talalay P (1994) Anticarcinogenic activities of sulforaphane and structurally related synthetic norbonyl isothiocyanates. Proc Natl Acad Sci USA 91:3147–3150. doi:10.1073/pnas.91.8.3147
27. Faig M, Bianchet MA, Talalay P, Chen S, Winski S, Ross D et al (2000) Structures of recombinant human and mouse NAD(P)H: quinone oxidoreductases: species comparison and structural changes with substrate binding and release. Proc Natl Acad Sci USA 97:3177–3182. doi:10.1073/pnas.050585797
28. Centner V, Massart DL, de Noord OE, de Jong S, Vandeginste BMV, Sterna C (1996) Elimination of uninformative variables for multivariate calibration. Anal Chem 68:3851–3858. doi:10.1021/ac960321m
29. Grohmann R, Schindler T (2008) Toward robust QSPR models: Synergistic utilization of robust regression and variable elimination. J Comput Chem 29:847–860. doi:10.1002/jcc.20831
30. Magdziarz T, Polanski J, Gieleciak R, Bak A (2008) Drug design toolbox http://prac.us.edu.pl/~zchorg/ddt Accessed 17 Jan 2008
31. Sybyl Computational Informatics Software for Molecular Modelers http://www.tripos.com/ Accessed 17 Jun 2008
32. CORINA Generation of 3D coordinates http://www.mol-net.com/software/corina/ Accessed 17 Jun 2008
33. Ghose A, Viswanadhan V, Wendoloski J (1998) Prediction of hydrophobic (Lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and CLOGP methods. J Phys Chem A 102:3762A–3772A. doi:10.1021/jp980230o
34. Audry E, Dubost JP, Colleter JC, Dallet P (1986) A new approach to structureactivity relations: the "Molecular Lipophilicity Potential". J Med Chem 21:71–72
35. Alhoniemi E, Himberg J, Parhankangas J, Vesanto J (2005) SOM Toolbox, Copyright (C) 2000–2005 by Esa Alhoniemi, Johan Himberg, Juha Parhankangas and Juha Vesanto http://www.cis.hut.fi/projects/somtoolbox/ Accessed 17 Jan 2008
36. Magdziarz T, Lozowicka B, Gieleciak R, Bak A, Polanski J, Chilmonczyk Z (2006) 3D QSAR study of hypolipidemic asarones by comparative molecular surface analysis. Bioorg Med Chem 14:1630–1643. doi:10.1016/j.bmc.2005.10.014
37. Gieleciak R, Magdziarz T, Bak A, Polanski J (2005) Modeling robust QSAR. 1. Coding molecules in 3D-QSAR - from a point to surface sectors and molecular volumes. J Chem Inf Model 45:1447–1455. doi:10.1021/ci0501488
38. Mickelson KE, Forsthoefel J, Westphal U (1981) Steroid-protein interactions. Human corticosteroid binding globulin: some physicochemical properties and binding specificity. Biochemistry 20:6211–6218. doi:10.1021/bi00524a047
39. Westphal U (1986) Steroid-protein interaction II. Springer, Berlin

40. Dunn WJ III, Wold S, Edlund U, Hellberg S, Gasteiger J (1984) Multivariate structure-activity relationships between data from a battery of biological tests and an ensemble of structure descriptors: the PLS method. Quant Struct Act Relat 3:131–137. doi:10.1002/qsar.19840030402
41. Wagener M, Sadowski J, Gasteiger J (1995) Autocorrelation of molecular surface properties for modeling corticosteroid binding globulin and cytosolic ah receptor activity by neural networks. J Am Chem Soc 117:7769–7775. doi:10.1021/ja00134a023
42. User and Reference Manual Quasar 4.0 http://www.biograf.ch Accessed Jan 2004
43. Robinson DD, Winn PJ, Lyne PD, Richards WG (1999) Self-organizing molecular field analysis: A tool for structure-activity studies. J Med Chem 42:573–583. doi:10.1021/jm9810607
44. Rucker C, Rucker G, Meringer M (2007) y-Randomization and its variants in QSPR/QSAR. J Chem Inf Model 47:2345–2357. doi:10.1021/ci700157b