Quantitative Structure–Cytotoxicity Relationships of Sesquiterpene Lactones derived from partial charge (Q)–based fractional Accessible Surface Area Descriptors (Q_frASAs)

ARTICLE *in* QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS · OCTOBER 2002

DOI: 10.1002/1521-3838(200208)21:3<276::AID-QSAR276>3.0.CO;2-S

2 AUTHORS, INCLUDING:

Thomas J. Schmidt
University of Münster
142 PUBLICATIONS **2,900** CITATIONS

SEE PROFILE

Quantitative Structure-Cytotoxicity Relationships of Sesquiterpene Lactones derived from partial charge (Q)-based fractional Accessible Surface Area Descriptors (Q_frASAs)

Thomas J. Schmidt[a]* and Jörg Heilmann[b]

[a] Institut für Pharmazeutische Biologie der Heinrich-Heine-Universität Düsseldorf, Universitätsstrasse 1, D-40225 Düsseldorf, Germany. E-mail: tj.schmidt@uni-duesseldorf.de
[b] Institut für Pharmazeutische Wissenschaften, ETH Zürich, Switzerland.

Dedicated to Prof. Dr. Irmgard Merfort, Institut für Pharmazeutische Biologie, Albert-Ludwigs-Universität Freiburg, on the occasion of her 50th birthday.

Abstract

In continuation of a previous QSAR study on the cytotoxicity of 20 sesquiterpene lactones (STLs) of the helenanolide type towards a mouse tumour cell line where a very strong correlation of activity with only two indicator variables encoding the nature of the present α,β-unsaturated carbonyl structure elements (cyclopentenone and α-methylene-γ-lactone structure) was found, it was the major goal of this study to establish a QSAR model for a set of STLs with wider structural variability. Cytotoxicity towards the human KB cervix carcinoma cell line was experimentally determined for a set of 40 STLs representing five different structural groups (2 germacranolides, 6 guaianolides, 23 pseudoguaianolides, 8 eudesmanolides and 1 carabranolide (cyclopropane type xanthanolide) and the resulting IC_{50} values were submitted to a QSAR study using the molecular modelling program MOE. As major result it could be shown that variance in STL cytotoxicity data can be explained to a high degree by electronic and surface properties. QSAR models of considerable internal and external predictivity could be obtained by PCR and PLS analysis of a descriptor set representing fractional accessible molecular surface areas (Q_frASAs). This set of descriptors is calculated by partitioning the molecular surface accessible to a spheric probe of radius 1.4 Å into fractions attributable to atoms within 14 charge intervals from −0.3 to 0.3 e. The applicability of such Q_frASA descriptors is validated by analysis of several sets of literature data, yielding QSAR models of good statistical quality. It is therefore assumed that Q_frASA descriptors may be of wider applicability in QSAR and QSPR.

* To receive all correspondence.

Key words: Natural Products, Sesquiterpene Lactones, Cytotoxicity, KB-cells, Principal components regression (PCR), Genetic algorithm-partial least squares (GA-PLS), partial charge-based fractional accessibe surface area descriptors (Q_frASA).

Abbreviations and symbols used:

STL:	Sesquiterpene lactone
PCR:	Principal component regression
PLS:	partial least squares statistics
GA-PLS:	genetic algorithm-driven variable selection using PLS
Q:	Partial atomic charges based on AM1 molecular orbital models
ASA:	Molecular surface area accessible to a speric probe of radius 1.4 Å
Q_frASA:	fractional accessible surface area descriptors based on Q and ASA
VSA:	VanderWaals surface area
PEOE:	Partial atomic charges based on "partitioning of orbital electronegativities"
PEOE_VSA:	fractional vanderWaals surface areas based on PEOE and VSA
r^2:	squared correlation coefficient
r_{pred}^2:	squared correlation coefficient of test set predictions
r^2_0:	squared correlation coefficient of test set regression trendlines forced through origin
s:	root-mean-square error in PCR, standard error of predicition in PLS
q^2:	squared correlation coefficient of leave-one-out predicitions
scv:	error of predicitions by leave-one-out cross validation, in PLS and PCR
n:	number of compounds
N:	number of significant principal components in PCR
LV:	latent variable (PLS-component)
k:	number of significant LVs in PLS

  0931-8771/02/0308-0276 $ 17.50+.50/0

1 Introduction

Cytotoxicity, as many other biological activities of sesquiterpene lactones (STLs), is known to be mediated by the presence of potentially alkylant structure elements capable of reacting covalently with biological nucleophiles, thereby inhibiting a variety of cellular functions [1] which directs the cells into apoptosis [2]. STLs have in the past been considered interesting leads to a new class of anti cancer agents (see lit. cited in [1]). Unfortunately, due to a low degree of selectivity towards tumour cells, it has not been possible to develop therapeutically useful anti cancer agents from this class of leads. It is known, that STLs affect the function of many enzyme systems and transcription factors so that their cytotoxicity is probably a consquence of interference with various target structures within the cell [1]. Despite the plethora of experimental studies found in literature on the cytotoxicity of particular STLs against many cell lines, little is known on the effects of different alkylant structure elements and of other structural factors on cytotoxicity in terms of quantitative structure-activity relationships (QSAR). This, however, would be an important step in the direction of rational lead optimization.

It can be assumed that any attempt to find QSAR relationships for STL cytotoxicity must take into account the capability of the molecules to engage in Michael-type addition to biological nucleophiles. This capability will for a large part depend on the presence of alkylant structure elements. This hypothesis was confirmed in our previous QSAR study on cytotoxicity of 20 helenanolide type STLs towards a murine Ehrlich ascites tumour cell line [3], where a very strong correlation with only two indicator variables encoding the nature of the present α,β-unsaturated carbonyl structure elements (cyclopentenone and α-methylene-γ-lactone structure) was found. Due to the high degree of structural similarity in the previous set of compounds, it appeared difficult to find further factors influencing activity, although significant contributions appeared to come from the number of H-bond acceptors and molecular conformation. It was therefore the primary goal of this study to establish a QSAR model for a set of STLs with a wider structural variability in order to gain a deeper insight into the biological function and pharmacological profile of this important class of plant secondary metabolites.

Secondly, – bearing in mind the ultimate goal of lead optimization – it would be desirable to have a QSAR model constructed from descriptors that are applicable to other types of STL bioactivity (such as their anti-inflammatory potential or their anti-protozoal activity). This would enable us to compare directly the impact of different structural properties on different activities, which would be an important breakthrough in the search for selective activity. It was hence a major objective of this study to find a global representation of STL molecular structure, based purely on theoretical molecular models (also important with respect to activity estimation of compounds not available for experimental measurements), that will in the future allow to construct such QSARs for different types of activity using the same series of descriptors.

2 Materials and Methods

2.1 Compounds

Sesquiterpene lactones **1** – **40** were obtained previously in our laboratories from different Asteraceae species. Purity of all compounds was shown to be $> 95\%$ by chromatographic analyses (GC, HPLC) and/or NMR.

2.2 Cytotoxicity Assay

The cytotoxicity of the compounds was determined using the KB cell line (ATCC CCL 17; HeLa cells). The tests were carried out in 96-well plates (Falcon) with an inoculum of 2.5×10^4 cells/ml. Test solutions were made as stocks in ethanol. Test concentrations were freshly prepared by diluting the stock solution with water to the required concentration. Final ethanol concentration was 1% (v/v) or less. Total assay volume was 150 µl. For quantification of the cytotoxicity, 15 µl of an aqueous solution of methylthiazolyltetrazolium chloride (MTT, Fluka, 5 mg/ml in PBS) was added after 72 h. During incubation at 37 °C for 4 h, the surviving cells metabolized MTT into an insoluble formazan dye. The culture medium was drawn off and the formazan dye was dissolved using 150 µl of 10% SDS (sodium dodecylsulfate) in water. After 24 h of incubation at room temperature, the optical density was measured at 540 nm using a microplate reader (MRX, Dynex Technologies). For determination of the IC_{50} values, the optical density was plotted against the log concentration and eight different concentrations were tested. Every test was performed in duplicates and all experiments were repeated at least twice. Maximum observed standard deviation was about 20% (absolute). Positive control measurements were performed with podophyllotoxin [4].

2.3 Molecular Modeling

For the helenanolide type compounds **1** – **23**, molecular models of a previous study were used [3]. All other compounds were modeled in an analogous way, using Hyperchem rel. 5.1. [5]. Starting geometries for each molecule (typically obtained by structural modification of compounds with known 3D-structure) were pre-minimised using the MM+ force field to an RMS gradient < 0.01. Each low-energy conformation was subsequently minimised to RMS gradient < 0.01 using the semi-empirical AM1 hamiltonian.

The compounds from the literature data sets [6, 7, 8] were modeled in an analogous way according to their depicted structures. The structures of the steroids from the Cramer data set [8] were corrected according to the respective Chemical Abstracts entries.

2.4 Descriptor Generation

QSAR descriptors were generated using the QuaSAR module implemented in the modelling package MOE [9]. PEOE_VSA descriptors [10] are calculated automatically using a 2D molecular graph and PEOE charges calculated by MOE. In order to obtain Q_frASA descriptors, the energy minimised geometries from Hyperchem were imported into a QuaSAR database along with partial charges obtained from the AM1 wavefunction. The accessible surface area (ASA) of an atom is calculated by the MOE program unit "qsurf.svl" as the part of a spheric surface area, obtained by adding the van der Waals radius and the probe radius (1.4 Å), that does not intersect any such sphere of any other atom in the molecule. This program unit was modified in such a way that the ASAs of atoms within 14 charge intervals of increment 0.05 e from $-0.3 \geq Q$ to $Q \geq +0.3$ e were added to yield the 14 Q_frASA descriptors.

2.5 Statistical Analysis, QSAR Model Construction

PCR was carried out with the QuaSAR module of MOE. Default settings were generally applied. Variables showing a relative importance < 0.1 (relative importance of a variable in a QuaSAR model is expressed as the absolute value of its normalized coefficient divided by the absolute value of the largest normalized coefficent in the model) were considered non-significant and eliminated from the analysis to obtain maximum leave-one-out q^2 values.

PLS and GA-PLS were carried out using the UNC QSAR server [11, 12, 13]. The maximum number of LVs was usually set to 5. In GA-PLS, the number of evolution steps ("crossovers") was chosen in such a way that no significant increase in fitness [11, 12, 13] was observed for more than 100 steps at the end of a calculation, which typically required about 1 000 crossovers. The chosen number of parents was set equal to the number of variables submitted. The number of cross validation groups was set equal to the number of compounds, i.e. a full leave-one-out cross validation was performed.

For a further validation of the model obtained by GA-PLS for 62 STLs (set 3 in Table 1), the data set was divided into seven different training- and test sets (42 and 20 compounds, respectively, in each case; see Table 3). This division was achieved by calculating molecular diversity using the QuaSAR fuction "diverse subset". Diverse subsets were calculated with respect to: 1. biological acitivty (pIC_{50}; ACT in Table 3), 2. all 14 frASA descriptors (DESC in Table 3), 3–7.: five different molecular fingerprints (MACCS structural keys, typed atom distances (TAD), typed atom triangles (TAT), typed graph distances (TGD), typed graph triangles (TGT) [9]). In each case, the 42 most diverse compounds were used as training set, leave-one-out cross validated PCR was carried out and the activities of the test set (remaining 20 compounds) were predicted on the basis of each individual training set model.

3 Results

Cytotoxicity towards the human KB cervix carcinoma cell line was experimentally determined for a set of 40 STLs representing five different structural groups (2 germacranolides, 6 guaianolides, 23 pseudoguaianolides, 8 eudesmanolides and 1 carabranolide (cyclopropane type xanthanolide)). The structures of the 40 compounds and the results of the cytotoxicity measurements (IC_{50} values (M) expressed in negative logarithmic form $= pIC_{50}$) are presented in Figure 1. Compounds **27**, **33** and **37** were essentially inactive ($pIC_{50} < 3$). These three compounds are the only ones in the data set without any α,β-unsaturated carbonyl groups. The fact that all of them possess an epoxy group indicates that this structure element on its own does not cause any significant cytotoxicity towards KB cells. The three compounds were therefore excluded from the set used for QSAR analysis.

In a previous study on the cytotoxicity of 20 helenanolide (= 10α-methylpseudo-guaianolide) type STLs against a murine Ehrlich ascites tumour cell line, it was demonstrated that cytotoxicity is to a major part explained by the presence or absence of different alkylant structure elements, a cyclopentenone- (CP) and/or an α-methylene-γlactone (MGL) structure. The QSAR relationship obtained by multiple linear regression (MLR), analysed in the form of Free-Wilson type indicator variables (1 in case of presence, 0 in case of absence of a respective structure element), explained approximately 90% of data variability on grounds of these two structure elements [3].

On these grounds, an analogous approach was tested for the 37 active compounds of the present study. Structure elements taken into account were α,β-unsaturated carbonyl structures (CP, MGL and α,β-unsaturated C5-carboxylic acid esters: tiglinate, angelate or sencionate C5ABU), as well as epoxide structures (EPO). As expected from the above mentioned result with compounds **27**, **33** and **37**, it was shown that only the CP, MGL and C5ABU descriptors yielded significant contributions to the resulting linear relationships obtained by multiple linear regression and by principal component regression (PCR, Eq. 1).

$$pIC_{50} = 0.756\ CP + 0.838\ MGL - 0.299\ C5ABU + 4.410 \quad (1)$$

$$r^2 = 0.698,\ s = 0.275;\ q^2 = 0.642,\ scv = 0.300$$

Relative importance of descriptors: CP: 1.0, MGL: 0.89, C5ABU: 0.37; (EPO: 0.03 = irrelevant, left out in Eq. 1); number of significant principal components = 3.

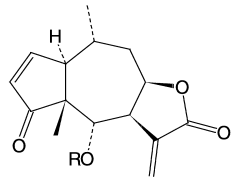
In spite of the lower degree of correlation, this result is in good agreement with our previous QSAR study [3]. It shows a clear dependence of activity on the presence/absence of the respective alkylant centers which explain about 70% of the variance in the biological data. However, further factors related to other molecular properties must be responsible for the unexplained variance.

It was therefore attempted to include further indicator variables, encoding the different substitution patterns and

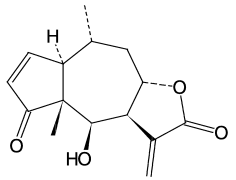
Table 1. Summary of PCR and PLS models obtained by GA-driven variable selection (GA-PLS) for 37 STLs of this study (set 1), for 27 STL from a literature data set [6] (set 2) and from the combined data sets (n = 62; two compounds were present in both data sets). All models are based exclusively on Q_frASA descriptors calculated as described in the text.

	Q_frASA														
Q	−7 ≤ − 0.30	−6 ≤ − 0.25.. > − 0.30	−5 ≤ − 0.20.. > − 0.25	−4 ≤ − 0.15.. > − 0.20	−3 ≤ − 0.10.. > − 0.15	−2 ≤ − 0.05.. > − 0.10	−1 ≤0.. > − 0.05	+1 >0.. <0.05	+2 ≥0.05.. <0.10	+3 ≥0.10.. <0.15	+4 ≥0.15.. <0.20	+5 ≥0.20.. <0.25	+6 ≥0.25.. <0.30	+7 ≥0.30	interc.
set 1 PCR		−0.0037	+0.0061		+0.0228			−0.0993			+0.0314		−0.0341		+4.4295
set 1* GA-PLS		−0.0047	+0.0052		+0.0203			−0.0991			+0.0324		−0.0389		+4.6158
n = 37; ***PCR:*** *N = 6; r^2 = 0.831; s = 0.206; q^2 = 0.724; scv = 0.266;* ***PLS:*** *k = 2; r = 0.828 ; s = 0.216; q^2 = 0.743 ; scv = 0.265; F = 82.1*															
set 2 PCR	−0.0136	−0.0067			+0.0172	+0.0192		−0.1124	+0.0041	+0.0078	+0.0104				+3.3450
set 2** GA-PLS		−0.0111			+0.0127	+0.0195	−0.0712	−0.0574			+0.0337		−0.0350		+5.1695
n = 27; ***PCR:*** *N = 9; r^2 = 0.907; s = 0.157; q^2 = 0.795; scv = 0.236;* ***PLS:*** *k = 5; r^2 = 0.858; s = 0.220; q^2 = 0.738; scv = 0.298; F = 25.3*															
set 1 + 2 PCR	−0.0052	−0.0109		−0.0059	+0.0165	+0.0193	−0.0298	−0.0799	+0.0029	+0.0053	+0.0360		−0.0559	−0.0150	+4.5624
set 1 + 2*** GA-PLS	−0.0052	−0.0072	+0.0057		+0.0143	+0.0144		−0.1033	+0.0019	+0.0029	+0.0329		−0.0706	−0.0124	+4.5171
n = 62; ***PCR:*** *N = 12; r^2 = 0.843; s = 0.200; q^2 = 0.749; scv = 0.256;* ***PLS:*** *k = 5; r^2 = 0.822; s = 0.225; q^2 = 0.725; scv = 0.280; F = 51.8*															

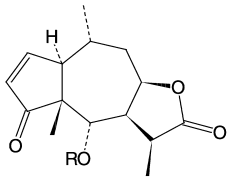
n = number of compounds; N = number of significant principal components; k = number of significant latent variables (PLS components).
* mod1 in Table 2
** mod2 in Table 2
*** mod3 in Table 2.



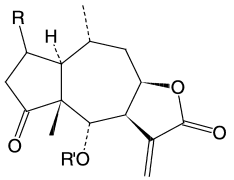
#	R	pIC_{50}
1	H	6.194 ± 0.057
2	ac	5.991 ± 0.063
3	*i*-but	6.076 ± 0.053
4	*i*-val	5.611 ± 0.063



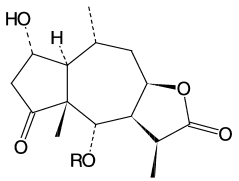
#	pIC_{50}
5	5.900± 0.036



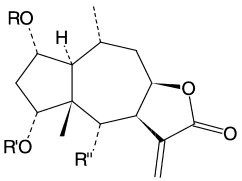
#	R	pIC_{50}
6	H	5.121± 0.017
7	ac	5.201± 0.021
8	i-val	5.489± 0.067



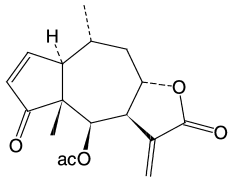
#	R	R‘	pIC_{50}
9	αOH	tig	5.254 ± 0.026
10	αOH	ang	5.211 ± 0.035
11	αOH	sen	4.909 ± 0.021
12	βOet	H	4.876 ± 0.030
13	βO*i*-pr	H	4.976 ± 0.049



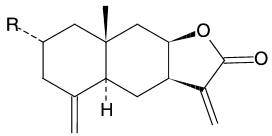
#	R	R‘	pIC_{50}
14	H	tig	3.951 ± 0.027
15	H	sen	3.959



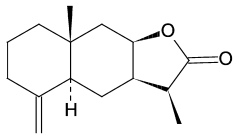
R	R‘	R‘‘	pIC_{50}
H	H	H	4.695 ± 0.015
ac	H	H	5.058 ± 0.019
ac	H	OH	5.642 ± 0.028
ac	H	Oac	5.827 ± 0.023
ac	ac	H	5.380 ± 0.083
H	H	Otig	4.691 ± 0.058
H	tig	OH	4.997 ± 0.081



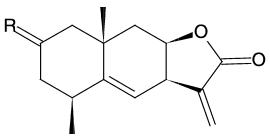
#	pIC_{50}
23	5.939 ± 0.004



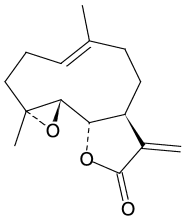
#	R	pIC_{50}
24	H	5.367 ± 0.056
25	OH	5.754 ± 0.017
26	Oac	5.421 ± 0.031



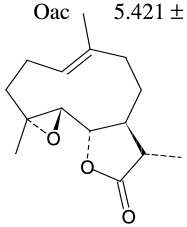
#	pIC_{50}
27	< 3



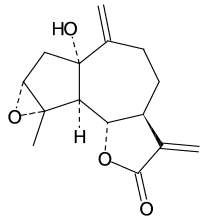
#	R	pIC_{50}
28	H_2	5.453 ± 0.094
29	O	5.424 ± 0.035



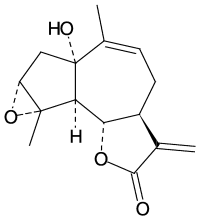
#	pIC_{50}
32	5.616 ± 0.031



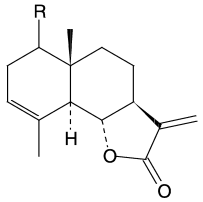
#	pIC_{50}
33	< 3



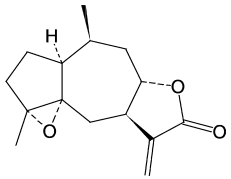
#	pIC_{50}
34	5.117 ± 0.030



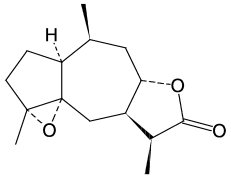
#	pIC_{50}
35	5.188 ± 0.049



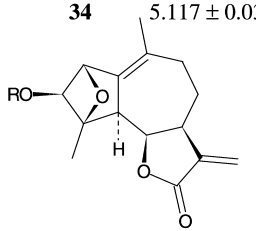
#	R	pIC_{50}
30	αOH	5.205 ± 0.046
31	βOH	5.090 ± 0.043



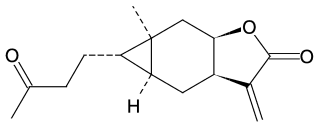
#	pIC_{50}
36	4.972 ± 0.085



#	pIC_{50}
37	< 3



#	R	pIC_{50}
38	H	4.765 ± 0.046
39	2-mebu	5.561 ± 0.043



#	pIC_{50}
40	4.829 ± 0.014

Figure 1. Structures of 40 STLs and their cytotoxicity towards KB cells ($-\log IC_{50}(M) = pIC_{50}$) experimentally determined for this study.

Table 2. Calculated and experimental pIC_{50} values for data set 1 (37 compounds of present study), set 2 (literature data for 27 STLs [6]) and combined data set 3, calculated from the GA-PLS models described in table 1.

compound-#	pIC_{50} (exp)	pIC_{50} (mod1)	pIC_{50} (mod2)	pIC_{50} (mod3)
1	6.194	5.997		5.892
2	5.991	5.791		5.704
3	6.076	5.884		5.972
4	5.611	5.872		5.966
5	5.900	5.880		5.581
6	5.121	5.450		5.474
7	5.201	5.029		4.971
8	5.490	5.067		5.176
9	5.254	5.020		5.123
10	5.211	4.901		5.140
11	4.909	4.854		4.882
12	4.876	5.096		5.171
13	4.976	5.109		5.208
14	3.951	4.253		4.363
15	3.959	4.166		4.200
16	4.695	4.813		4.547
17	5.058	5.228		5.035
18	5.642	5.213		4.999
19	5.827	6.004		5.852
20	5.380	5.346		5.145
21	4.691	4.738		4.695
22	4.997	5.226		5.285
23	5.939	6.188		6.053
24	5.367	5.247		5.318
25	5.755	5.806		5.791
26	5.421	5.304		5.260
28	5.454	5.403		5.487
29	5.424	5.077		5.148
30	5.205	5.227		5.129
31	5.090	5.234		5.143
32	5.616	5.852		5.670
34	5.117	5.229		5.213
35	5.188	5.120		5.193
36	4.972	5.011		4.937
38	4.765	4.800		4.692
39	5.561	5.275		5.382
40	4.829	4.999		4.953
42	4.582		4.775	5.208
43	5.186		5.182	5.223
44	5.071		5.172	5.341
45	5.914		5.622	5.759
46	5.164		5.219	5.184
47	4.198		4.189	3.974
49	6.276		6.367	6.298
50	6.456		6.432	6.399
51	5.903		5.520	5.691
52	5.609		5.596	5.546
53	5.294		5.522	5.330
54	4.979		5.482	5.331
55	5.213		5.155	5.104
56	5.406		5.408	5.642
57	5.684		5.545	5.536
58	5.260		4.977	5.234
59	5.000		4.820	5.127
60	5.060		5.027	4.980
61	5.886		5.891	5.850
62	5.738		5.769	5.994
63	5.860		5.961	5.818
64*	6.120		5.961	*
65*	5.900		5.862	*
66	5.492		5.500	5.456
67	5.600		5.323	5.464
69	4.943		5.267	5.011
70	4.943		5.193	5.174

* identical with compound **1** and compound **5**, therefore included only once in mod3.

Table 3. Test set compounds (numbering see Figures 1 and 2) used for external validation of model mod3 (see Tables 1 and 2). Division into training set (n = 42) and test set (n = 20) was carried out as described in the Materials and Methods section. r_{pred}^2 and r^2_0 and slope values were obtained from regression trendlines in plots of pIC_{50} vs. predicted pIC_{50} in each case.

	ACT	DESC	MACCS	TAD	TAT	TGD	TGT
	4	**2**	**2**	**3**	**2**	**2**	**2**
	7	**4**	**4**	**5**	**3**	**4**	**5**
	8	**5**	**5**	**6**	**5**	**5**	**6**
	10	**8**	**9**	**7**	**6**	**6**	**9**
	13	**10**	**10**	**8**	**7**	**9**	**10**
	15	**13**	**11**	**9**	**9**	**10**	**11**
	17	**17**	**13**	**10**	**10**	**11**	**15**
	20	**19**	**14**	**11**	**12**	**13**	**18**
	21	**24**	**18**	**14**	**14**	**14**	**23**
	22	**26**	**19**	**20**	**15**	**23**	**24**
	26	**30**	**22**	**21**	**17**	**24**	**28**
	31	**31**	**23**	**28**	**23**	**26**	**31**
	34	**32**	**31**	**29**	**24**	**28**	**43**
	43	**34**	**42**	**31**	**26**	**31**	**46**
	51	**36**	**44**	**42**	**31**	**35**	**53**
	52	**43**	**46**	**46**	**43**	**43**	**58**
	54	**54**	**60**	**49**	**44**	**46**	**59**
	58	**55**	**66**	**58**	**58**	**53**	**60**
	60	**58**	**69**	**61**	**67**	**54**	**61**
	70	**63**	**70**	**66**	**70**	**58**	**63**
r_{pred}^2	0.688	0.643	0.623	0.738	0.833	0.711	0.691
slope	0.898	0.925	0.955	0.954	1.375	1.016	0.986
r^2_0	0.679	0.639	0.622	0.735	0.770	0.711	0.690
slope	0.999	1.002	0.988	1.015	0.997	0.990	1.021

skeletal types in the analysis. This attempt was unsuccessful. Moreover, attempts to combine these indicator variables with a variety of "classical" QSAR descriptors such as logP, hydration energy, molecular refractivity, polarizability, HOMO- and LUMO energies, dipole moments, etc. as calculated by Hyperchem/Chem+ [5] did not lead to satisfactory results.

To search for descriptors of molecular structure relevant in explaining the cytotoxic activity, the program package MOE (Molecular Operations Environment, [9]) was used. The MOE QuaSAR module automatically generates several hundred molecular descriptors from 2D and 3D molecular information [9]. Statistical analysis to create QSAR models is based on principal component regression (PCR).

Search for relevant descriptors within the descriptor set automatically calculated by QuaSAR revealed a considerable degree of correlation with fractional van der Waals surface area descriptors (PEOE_VSA descriptors [10]). This subset of QuaSAR descriptors is based on atomic partial charges (PEOE charges obtained by the Method of Gasteiger et al. according to [9]) and each atom's contribution to the molecule's van der Waals (vdW) surface. Each PEOE_VSA represents the fraction of a molecule's vdW surface area attributable to atoms within a certain range of partial charge. The total vdW surface is subdivided into 14 charge intervals (7 negative and 7 positive, increments of 0.05 from +0.3 to −0.3 e). Correlation of the total 37 compounds with the PEOE_VSAs (of the 14 descriptors only 9 were relevant) yielded a PCR model with correlation coefficients of $r^2 = 0.77$, $q^2 = 0.55$. This indicated that STL cytotoxicity may be expressed as a function of atomic partial charges and their distribution on the molecular surface.

In our previous QSAR study [3] a significant influence of molecular conformation on cytotoxicity was found. The PEOE_VSA descriptors, however, do not contain any stereochemical information, since they are calculated from 2D molecular graphs. The influence of various molecular shape descriptors from the MOE/QuaSAR set was therefore tested by combining them with the relevant PEOE_VSAs. No significant improvement of the model could be achieved in this way.

Since it could be expected that partial atomic charges obtained by quantum mechanical calculations would take into account effects of delocalization in a more realistic way than PEOE charges, and in order to include information on the molecules' steric properties, a new set of descriptors was created, encoding partial charge (*Q*)- based *fr*actional *a*ccessible *s*urface *a*reas, termed "Q_frASA". Q_frASAs are calculated from 3D molecular models and based on the molecular surface accessible to a spherical probe (radius 1.4 Å) and partial charges (Q) resulting from the semiempirical molecular orbital method AM1. A total of 14 frASA descriptors were calculated (charge intervals of 0.05 from $-0.30 \geq Q$ to $Q \geq +0.30$ e).

This set of descriptors thus differs from the PEOE_VSAs in two respects, firstly it takes into account a more "realistic" surface of the molecules with respect to interaction with the biological target and depending on their 3D shape, and secondly, it uses a partial charge model based on quantum chemical calculations and should thus model more accurately the effects of electron delocalization (important for electrophilicity of the alkylant structure elements).

When all 14 Q_frASA descriptors were correlated with activity, the result was clearly superior to that obtained with the PEOE_VSAs. The PCR yielded significant contributions from 6 of the 14 descriptors. The correlation coefficients were $r^2 = 0.831$ and $q^2 = 0.724$ (leave-one-out cross validation) using the 6 principal components resulting from these descriptors (see Table 1). The set of Q_frASAs was thereby shown to yield a result superior to that obtained with the PEOE_VSAs.

In addition to PCR, the set of Q_frASA descriptors was subjected to another method of model building, genetic algorithm-driven variable selection in combination with PLS analysis (GA-PLS [11, 12, 13]). In this method, a family of most significant variable combinations (in terms of a fitness criterion related to q^2) is found by applying the principles of evolution to the data matrix [11, 12, 13].

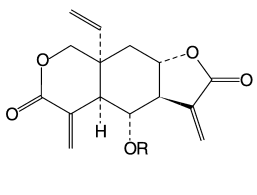
The best variable combination found after 1 000 evolution steps ("crossovers") on all 14Q_frASAs yielded a model consisting of 2 significant PLS components (latent variables, LVs) constructed from the same 6 variables as found by PCR, and showing correlation coefficients of $r^2 = 0.828$ and $q^2 = 0.743$ (Table 1).

The application of Q_frASA descriptors should lead to reasonable results also for other related data sets. In order to test the applicability with other types of STLs, the data set of Kupchan et al. was chosen, where a variety of STLs, also from structural groups not covered by the present set of compounds, was tested for cytotoxicity against the KB cell line [6]. 30 compounds of this data set (Figure 2) were modeled and treated in essentially the same way as described above. When the biological data were correlated with the 14 Q_frASAs by PCR, the initial result with all descriptors did show a reasonable $r^2 = 0.77$ but the q^2 value was only 0.13. Successive elimination of descriptors with low relative importance, aiming at a maximisation of q^2, led to a model with 10 descriptors and $r^2 = 0.77$ and $q^2 = 0.46$. Inspection of the leave-one-out predictions showed that three compounds (**41**, **48** and **68**) represented obvious outliers. When these compounds were omitted, the result was dramatically improved although the number of descriptors could be reduced to 9 ($r^2 = 0.91$, $q^2 = 0.79$; $n = 27$; see Table 1). A very similar result was obtained by submitting the reduced set of compounds ($n = 27$) to GA-PLS (1 000 crossover steps). The best PLS model thus obtained consisted of five LVs, constructed from 7 descriptors ($r^2 = 0.86$, $q^2 = 0.74$; see Table 1), and is thus almost equivalent to the PCR model.

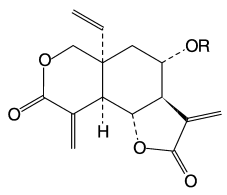
The fact that five of the descriptors were identified as important for cytotoxicity in both models, each of them showing the same sign of its PCR and PLS coefficients (see Table 1) encouraged us to attempt at finding a common QSAR for both data sets representing cytotoxicity data of STLs towards the same cell line.

When the two data sets were united and the data for 62 STLs (37 from the present study and 25 from the literature set, since two compounds (**1**≡**64** and **5**≡**65**) were present in both sets) were correlated by PCR with the 14 descriptors, a model resulted which contained significant contributions from 12 variables (12 principal components used) and showing $r^2 = 0.84$, $q^2 = 0.75$ ($n = 62$). GA-PLS (500 steps) led to a model with $r^2 = 0.82$ and $q^2 = 0.73$, made up of 5 LVs constructed from 11 variables (Table 1). The experimental pIC_{50} values are plotted against those calculated by this model in Figure 3.

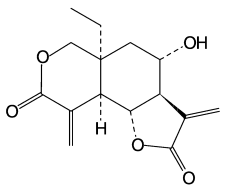
It has recently been argued that a high q^2 alone is not a sufficient criterion to estimate the predictive capacity of a QSAR model and additional criteria have been proposed in



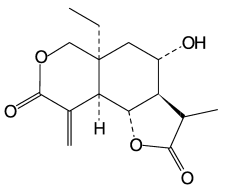
#	R	pIC_{50}
41	H	3.896
42	ac	4.582



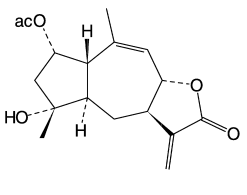
#	R	pIC_{50}
43	H	5.186
44	ac	5.071
45	meac	5.914



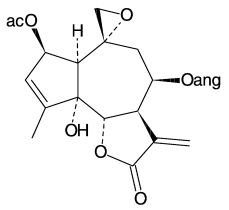
#	pIC_{50}
46	5.164



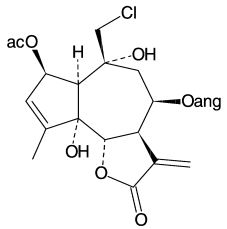
#	pIC_{50}
47	4.198



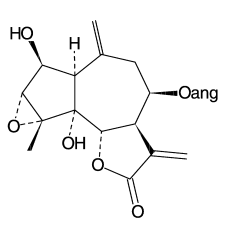
#	pIC_{50}
48	5.124



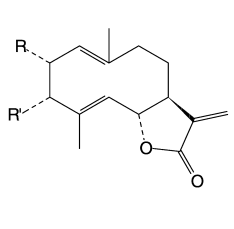
#	pIC_{50}
49	6.276



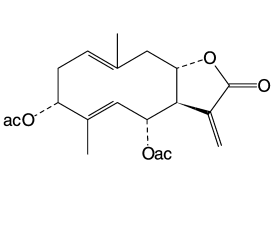
#	pIC_{50}
50	6.456



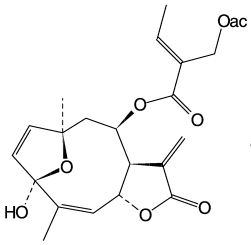
#	pIC_{50}
51	5.903



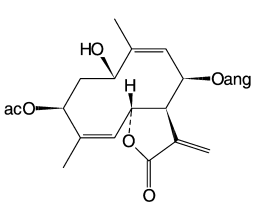
#	R	R'	pIC_{50}
52	H	H	5.609
53	OH	H	5.294
54	H	OH	4.979



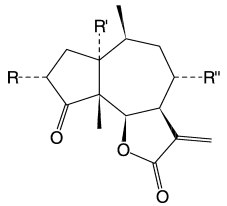
#	pIC_{50}
55	5.213



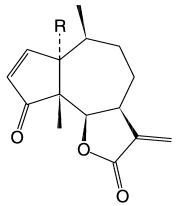
#	pIC_{50}
56	5.406



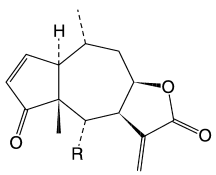
#	pIC_{50}
57	5.684



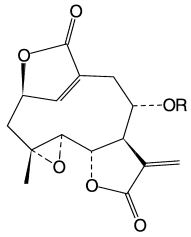
#	R	R'	R''	pIC_{50}
58	H	OH	H	5.260
59	OH	H	H	5.000
60	H	H	OH	5.060



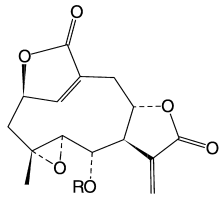
#	R	pIC_{50}
61	OH	5.886
62	H	5.738



#	R	pIC_{50}
63	H	5.860
64	OH*	6.120
65	βOH, 8α**	5.900



#	R	pIC_{50}
66	meac	5.492
67	sen	5.600



#	R	pIC_{50}
68	H	3.909
69	ac	4.943
70	*i*-but	4.943

* identical with compound **1** of present data set (fig. 1)
** identical with compound **5** of present data set (fig. 1)

Figure 2. Structures of 30 STLs and their cytotoxicity towards KB cells ($-\log IC_{50}(M) = pIC_{50}$) taken from literature [6].

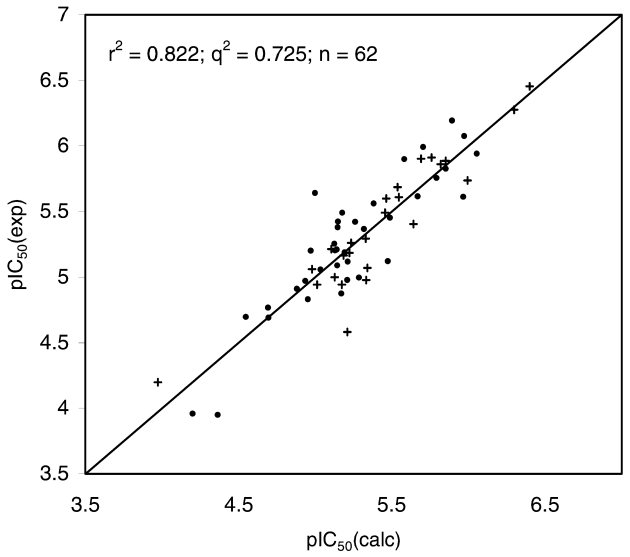


Figure 3. Experimental vs. calculated pIC_{50} values from GA-PLS model for 62 STLs combined from data of this study (n = 37, ●) and a data set from literature [6] (n = 25, +; see text).

order to avoid unpredicitve models with high q^2 [14]. This 12-variable model was therefore further validated by subdividing the 62 compounds into several training sets (42 compounds in each case) and test sets (20 compounds in each case). This division was performed by calculating molecular diversity using the method implemented in MOE (QuaSAR) and using seven different criteria: biological activity (pIC_{50}), the set of Q_frASA descriptors, and five different molecular fingerprints. In each case the 42 most diverse compounds were used as training set and the remaining 20 compounds which were not used in the model building process were used as test set (see Table 3). For each training set, PCR was carried out independently, yielding an average $r^2 = 0.85$ and $q^2 = 0.72$. For each test set, predictions were made on the basis of the model obtained with the corresponding training set. The average test set prediction yielded $r_{pred}^2 = 0.70$ (min = 0.62; max = 0.83). Furthermore, as suggested by Tropsha and coworkers [14], the regression trendlines in plots of predicted vs. experimental pIC_{50}s and vice versa were investigated with respect to deviations of their slopes from 1 and their intercepts from 0. In all cases, at least one trendline forced through the origin showed an r^2_0 almost as high as r_{pred}^2 (average $r_{pred}^2 = 0.70$; average $r^2_0 = 0.69$) and in all cases, the slope was close to 1 (max = 1.021, min = 0.988). The model thus satisfies stringent validation criteria for predictive quality, showing that this combination of Q_frASA descriptors yields also considerable external predictivity and that the model thus obtained is fairly robust and stable.

The models obtained for STL cytotoxicity using Q_frASA descriptors indicate that these descriptors are capable of covering a major part of molecular information relevant for the ability of a group of chemicals to interact with their cellular targets. In order to test the applicability of this type of descriptors to other QSAR problems, they were applied to two literature data sets for non-STL chemicals.

Firstly, a data set on fish toxicity of a series of 18 Michael acceptors (subset 2, i.e. compounds 36–53, of the QSAR study of Karabunarliev et al. [7] on toxicity of 100 diverse chemicals towards the fathead minnow (*Pimephales promelas*) was chosen. In this case it could be anticipated that the same basic mechanism, i.e. alkylation of biological nucleophiles, underlies the biological activity as in case of the STLs.

The Michael acceptors investigated by Karabunarliev et al. comprise a variety of acrylates, methacrylates and some other α,β-unsaturated carbonyl structures. When the pLC_{50} values were correlated by PCR with Q_frASA descriptors obtained in the same way as for the STLs, a correlation of $r^2 = 0.84$, $q^2 = 0.61$ was obtained with six of the descriptors (six significant components; see Table 4). After 1 000 evolution steps using the GA-PLS method, a combination of five descriptors was found, yielding a model with four significant latent variables and $r^2 = 0.81$ and $q^2 = 0.53$ (see Table 4). This result, although less satisfactory than that obtained with the STLs, demonstrates the applicability of the presented method to toxicity QSAR of α,β-unsaturated carbonyl structures other than STLs. The lower statistical quality in this case is most certainly due to the fact that the biological data represent toxicity towards whole fish rather than one particular cell type.

Finally, in order to investigate whether Q_frASA descriptors may also be applied in cases where biological acitivity is dependent entirely on specific non-covalent binding, the well known steroid data set used by Cramer et al. in the initial work on CoMFA [8] was analysed. The binding data of 31 steroids for corticosteroid binding globulin (CBG) and for 21 of these compounds concerning testosterone binding globulin (TeBG) were correlated with Q_frASA descriptors. In the case of CBG binding for the 31 compounds, PCR using all 14 frASAs yielded $r^2 = 0.91$ and $q^2 = 0.69$. Elimination of the non-significant Q_frASAs led to $r^2 = 0.91$ and $q^2 = 0.77$ with the remaining 10 descriptors. Application of GA-PLS (500 evolution steps) led to a model containing two LVs, with contributions from 7 descriptors and correlation coefficients of $r^2 = 0.85$, $q^2 = 0.78$ (see Table 5).

In the same manner, the binding data for 21 compounds to TeBG were analysed. PCR of all Q_frASAs led to $r^2 = 0.94$ and $q^2 = 0.65$ which could be improved by eliminating 7 insignificant descriptors which yielded a model with $r^2 = 0.93$ and $q^2 = 0.83$. GA-PLS (500 steps) led to selection of 6 variables, combined in 5 significant LVs and showing $r^2 = 0.91$, $q^2 = 0.80$ (see Table 5; since the number of 5 LVs included in a model for 21 compounds is quite high, this calculation was also carried out allowing for a maximum of 3 LVs. This led to a very similar model, showing $r^2 = 0.84$, $q^2 = 0.72$).

4 Discussion

Before a discussion of the QSAR model for STLs obtained in this study, it is necessary to focus on the type of information encoded in the Q_frASA descriptors applied here for the first time. Clearly, the total set of Q_frASAs represents a global description of steric properties (acces-

Table 4. Summary of PCR and PLS models obtained by GA-driven variable selection (GA-PLS) for a literature data set [7] on fish toxicity of 18 α,β-unsaturated carbonyl compounds, based on Q_frASA descriptors. Biological data (pLC_{50} [M]) vary between 2.6 and 6.5.

	Q_frASA														interc.
Q	−7 ≤ −0.30	−6 ≤ −0.25.. > −0.30	−5 ≤ −0.20.. > −0.25	−4 ≤ −0.15.. > −0.20	−3 ≤ −0.10.. > −0.15	−2 ≤ −0.05.. > −0.10	−1 ≤0.. > −0.05	+1 >0.. <0.05	+2	+3 ≥0.05.. <0.10	+4 ≥0.10.. <0.15	+5 ≥0.15.. <0.20	+6 ≥0.20.. <0.25	+7 ≥0.25.. <0.30	≥0.30
set 3 PCR	−0.0243	+0.0316				−0.0563	−0.0266	+0.0729			+0.0514				+4.5568
set 3 GA-PLS	−0.0278	+0.0311				−0.0547		+0.0672			+0.0502				+4.4817
n = 18; ***PCR:*** *N = 6; r^2 = 0.841; s = 0.428; q^2 = 0.609; scv = 0.739;* ***PLS:*** *k = 4; r^2 = 0.810; s = 0.551; q^2 = 0.532; scv = 0.865; F = 13.9*															

n = number of compounds; N = number of significant principal components; k = number of significant latent variables (PLS components).

Table 5. Summary of PCR and PLS models obtained by GA-driven variable selection (GA-PLS) for a literature data set [8] on binding of steroids to CBG (31 compounds, set 4A) and TeBG (21 compounds, set 4B) based on Q_frASA descriptors.
Biological data (pK [M]) vary between 5.0 and 7.9 (CBG) and between 5.3 and 9.7 (TeBG).

	Q_frASA														interc.
Q	−7 ≤ −0.30	−6 ≤ −0.25.. > −0.30	−5 ≤ −0.20.. > −0.25	−4 ≤ −0.15.. > −0.20	−3 ≤ −0.10.. > −0.15	−2 ≤ −0.05.. > −0.10	−1 ≤0.. > −0.05	+1 >0.. <0.05	+2 ≥0.05.. <0.10	+3 ≥0.10.. <0.15	+4 ≥0.15.. <0.20	+5 ≥0.20.. <0.25	+6 ≥0.25.. <0.30	+7 ≥0.30	
set 4A PCR	+0.0550	+0.0350		+0.0305		+0.0682			+0.0080	−0.0051	−0.0146	−0.0190	+0.0756	+0.0872	+0.2541
set 4A GA-PLS	+0.0162		+0.0152	−0.0188	−0.0379	+0.0390	−0.1118						+0.0750		+5.7969
n = 31; ***PCR:*** *N = 10; r^2 = 0.905; s = 0.327; q^2 = 0.773; scv = 0.514;* ***PLS:*** *k = 2; r^2 = 0.849; s = 0.435; q^2 = 0.781; scv = 0.523; F = 78.5*															
set 4B PCR	−0.0700	−0.0655	+0.0392		+0.1307			−0.5453				+0.0150	+0.1769		+8.5990
set 4B GA-PLS	−0.0537	−0.0439			+0.1670		+0.0496	−0.7908					+0.1179		+8.3237
n = 21; ***PCR:*** *N = 7; r^2 = 0.932; s = 0.307; q^2 = 0.831; scv = 0.490;* ***PLS:*** *k = 5; r^2 = 0.908; s = 0.424; q^2 = 0.797; scv = 0.629; F = 29.5*															

n = number of compounds; N = number of significant principal components; k = number of significant latent variables (PLS components).

sible surface area) and electronic features (partial charges) for each molecule. By mapping the atomic charges obtained from quantum mechanical calculations onto the accessible surface, information on reactivity, i.e. the capacity to engage in covalent as well as non-covalent interactions, is included. By dividing the molecular accessible surface into fractions corresponding to regions of different charge, information on the overall distribution of negative and positive charge on the surface is inherent in the total set of descriptors. Thus, the entire set of Q_frASA descriptors can be considered as a "holistic" representation of each molecule's possibilities to engage in all types of interactions with a biological system. In an analogous way as previously stated for PEOE_VSA descriptors [10], Q_frASAs are therefore a means of representing the structural information relevant for biological activity in terms of a coherent "chemistry space" (both approaches actually represent a generalization of the polar surface area (PSA) descriptors which have become increasingly popular in QSAR/QSPR research, see e.g. [15]), which could be applicable in cases where it is desirable to investigate and compare QSAR and/or QSPR for a given set of chemicals with respect to different activities or properties. As a major difference from the PEOE_VSA descriptors, Q_frASA descriptors inherently contain 3D information since both, the accessible surface as well as the atomic charges used, depend on the 3D shape of the molecules.

It is clear that certain interactions (and thus, certain parts of the surface/charge continuum) are of higher relevance to a particular biological activity than others. In this respect, either the identification of relevant descriptors by PCR or the genetic algorithm-driven variable selection coupled to PLS serve the purpose of identifying those significant regions and to eliminate insignificant ones (i.e. "noise"). It can be expected that the Q_frASA approach will make it possible to describe and explain structure-activity relationships for different biological activities of a given set of compounds in the same context, i.e. using identical descriptors and evaluating the different influences of these descriptors on different activities.

By adding up the surface fractions corresponding to each charge interval into one larger patch, information on the size and location of each individual patch is lost. The fact that Q_frASAs perform quite well might indicate that such information is not needed to explain and predict the biological activity, which, on the other hand, could be interpreted in such a way that the interactions under consideration do not require any specific orientation of the relevant surface regions relative to each other. On the other hand, as shown for the steroid data set, where specific interaction with the proteins' binding sites is certainly important, Q_frASAs may also be of relevance in treating such activities that depend on a specific geometric arrangement of binding points within the structure. This observation might be explained by the fact that the structures of the steroids are relatively similar on the whole and that the relevant Q_frASAs corresponding to the interaction sites are "coincidentally" located in similar regions. Investigations on data sets for specific receptor binding with a larger overall structural diversity (i.e. non-congeneric series of drugs) will have to show whether Q_frASAs can also be applied in such cases.

One advantage of Q_frASA-based QSAR is the possibility to visualize the contributions of each variable to activity by color-mapping the respective PLS (or PCR) coefficients on the molecules' accessible surface. Representative structures of four STLs are shown in Figure 4.Regions with positive contribution to activity (i.e. positive sign of PLS coefficient) are shown in blue, regions with a detrimental effect on activity (negative coefficients) in red. The most prominent blue patches correspond to the cyclopentenone and methylene lactone structures. This expected result corresponds to that obtained with the simple equation 1 where the two structure elements are represented by indicator variables. On the other hand, it contains elements of accessibility and of mesomeric effects so that the impact of molecular geometry and electronic structure is incorpo-

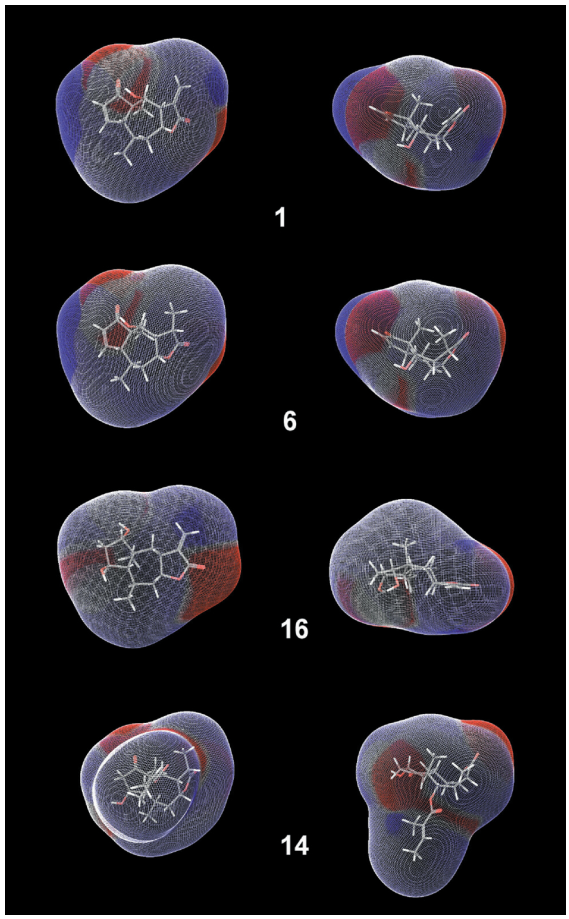


Figure 4. PLS coefficients from GA-PLS model for 62 STLs (see Table 1 and Figure 3) mapped onto the accessible surface of STLs **1**, **6**, **16** and **14** (shown from two different viewing angles in each case; structures see Figure 1).
Blue color corresponds to a positive, red to a negative coefficient (i.e. enhancing and reducing influence on cytotoxicity, respectively). For discussion see text.

rated and the contributions of the alkylant centers are thus encoded in a more subtle way.

The regions with most prominent negative contributions can be attributed to carbonyl oxygen- and carbon atoms, in the lactone and cyclopentenone structures. This might be related to the reactivity of the β-carbon in these α,β-unsaturated carbonyl structures being decreased in cases where the carbonyl carbon is too easily accessible.

Further influences are less straightforward to discuss and are apparently of a modulatory nature. It is important to bear in mind the nature of the descriptor set as a "holistic" description of molecular shape and electronic properties which – in spite of making it difficult to discuss in detail the influence of isolated structure elements – leads to a high explanatory and predicitive quality of the models. The fact that Q_frASAs encode the molecules' overall capacity to interact with any kind of biological target is quite likely to result in a wide range of applications to QSAR and QSPR problems, in a similar context as previously stated for PEOE_VSA descriptors [10].

Further work to evaluate the use of Q_frASAs is currently directed towards NF-κB inhibitory activity of STLs [16]. Moreover, studies on the different influence of the Q_frASAs on the cytotoxic activity towards different cell lines, in order to find a basis for selectivity towards certain tumors, is presently in progress. Important points that need further evaluation are the size of the charge interval (i.e. the number of Q_frASA descriptors used) and its possible adaptation to the specific requirements for a particular set of compounds, as well as the possibility to preserve the "lost information" on steric orientation and size of the Q_frASA "patches", which might be achieved by using a "WHIM"-type approach [17] or combination with 3D-molecular fingerprinting. Moreover, it will be of interest to investigate frASAs based on properties other than partial atomic charges (e.g. MO-coefficients, reactivity indices [18], atomic superdelocalizability [7, 19], and others).

Acknowledgements

Thanks are due to Dr. Chris Williams, Chemical Computing Group Inc., Montreal, Canada and Dr. Alexander Tropsha, Laboratory for Molecular Modeling, School of Pharmacy, University of North Carolina at Chapel Hill, USA, for critically reviewing the manuscript and for valuable hints and discussions. We thank Mr. Michael Wasescha (Institut für Pharmazeutische Wissenschaften, ETH Zürich) for excellent assistance in performing the cytotoxicity assay.

References

[1] Schmidt, T. J., Toxic Activities of Sesquiterpene Lactones - Structural and Biochemical Aspects, *Current Org. Chem. 3*, 577 – 605 (1999).

[2] Dirsch, V. M., Stuppner, H., Vollmar, A. M., Helenalin triggers a CD95 death receptor-independent apoptosis that is not affected by overexpression of Bcl-xL or Bcl-2, *Cancer Res. 61*, 5817 – 5823 (2001).

[3] Schmidt, T. J. Quantitative structure-cytotoxicity relationships within a series of helenanolide type sesquiterpene lactones. (Helenanolide Type Sesquiterpene Lactones, IV.), *Pharm. Pharmacol. Lett. 9*, 9 – 13 (1999).

[4] Heilmann, J., Wasescha, M. R., Schmidt, T. J., The influence of glutathione and cysteine levels on the cytotoxicity of helenanolide type sesquiterpene lactones against KB cells, *Bioorg. Med. Chem. 9*, 2189 – 2194 (2001).

[5] Hyperchem v. 5.1 is a product of Hypercube Inc., http://www.hyper.com/

[6] Kupchan, S. M., Eakin, M. A., Thomas, A. M. Tumor Inhibitors. 69. Structure-Cytotoxicity Relationships among the Sesquiterpene Lactones, *J. Med. Chem. 14*, 1147 – 1152 (1971).

[7] Karabunarliev, S., Mekenyan, O. G., Karcher, W., Russom, C. L., Bradbury, S. P., Quantum-chemical Desriptors for Estimating the Acute Toxicity of Electrophiles to the Fathead minnow (*Pimephales promelas*): An Analysis Based on Molecular Mechanisms, *Quant. Struct.-Act. Relat. 15*, 302 – 310 (1996).

[8] Cramer, R. D. III, Patterson, D. E., Bunce, J. D., Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins, *J. Am. Chem. Soc. 110*, 5959 – 5967 (1988).

[9] MOE (the Molecular Operating Environment) Version 2001.01; available from Chemical Computing Group, 1010 Sherbrooke Street West, Suite 910, Montreal, Quebec, Canada H3A 2R7 *http://www.chemcomp.com/;* for documentation of QuaSAR descriptors see *http://www.chemcomp.com/feature/descr.htm*.

[10] Labute, P., A Widely Applicable Set of Descriptors, *J. Mol. Graphics Mod. 18*, 464 – 477 (2000); see also *Journal of the Chemical Computing Group*, *http://www.chemcomp.com/feature/vsadesc.htm*.

[11] UNC QSAR server: *http://mmlin1.pha.unc.edu/~jin/QSAR/*

[12] Cho, S. J., Cummins, D., Bentley, J., Andrews, C. W., Tropsha, A., "Back" to 2D QSAR: Application of Genetic Algorithms and Partial Least Squares to Variable Selection of Topological Indices, *http://mmlin1.pha.unc.edu/~jin/QSAR/ GA_PLS/gapls.html*

[13] Cho S. J., Zheng, W., Tropsha, A. Rational Combinatorial Library Design. 2. Rational Design of Targeted Combinatorial Peptide Libraries Using Chemical Similarity Probe and the Inverse QSAR Approaches. *J. Chem. Inf. Comput. Sci. 38*, 259 – 276 (1998).

[14] Golbraikh, A., Tropsha, A., Beware of q^2!, *J. Mol. Graphics Mod. 20*, 269 – 276 (2002).

[15] Ertl, P., Rohde, B., Selzer, P., Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties, *J. Med. Chem. 43*, 3714 – 3717 (2000).

[16] Rüngeler, P., Castro, V., Mora, G., Gören, N., Vichnewski, W., Pahl, H. L., Merfort, I., Schmidt, T. J., Inhibition of Transcription Factor NF – κB by Sesquiterpene Lactones – A Proposed Molecular Mechanism of Action, *Bioorg. Med. Chem. 7*, 2343 – 2352 (1999).

[17] Todeschini, R., Lasagni, M., Marengo, E., New Molecular Descriptors for 2D and 3D Structures. Theory, *J. Chemom. 8*, 263 – 272 (1994).

[18] Fleming, I. *Frontier Orbitals and Organic Chemical Reactions*, Wiley and Sons, London 1976.

[19] Schüürmann, G., Quantitative Structure-Property Relationships for the Polarizability, Solvatochromic Parameters and Lipophilicity, *Quant. Struct.-Act. Relat. 9*, 326 – 333 (1990).

Received on February 25, 2002; accepted on April 9, 2002