Monatsh Chem (2010) 141:1321–1328
DOI 10.1007/s00706-010-0405-2

ORIGINAL PAPER

A quantitative structure–activity relationship study of the skin-irritant effect of thietanes

Issa Kakoie Dinaki · Morteza Zarrineh

Received: 2 March 2010 / Accepted: 14 September 2010 / Published online: 15 October 2010

Abstract Quantitative structure–activity relationships (QSAR) have been used to study the skin-irritant effect of 84 thietanes. A suitable set of molecular descriptors was calculated and the important descriptors were selected with the aid of the genetic algorithm and stepwise multiple regression methods. On the basis of principal-components analysis the data set was divided into 67 molecules in a training set and 17 molecules in a test set. The models were validated by use of leave-one-out cross-validation, an external test set, and a Y-randomization test. Comparison of the results obtained indicated the superiority of the genetic algorithm over stepwise multiple regression for feature selection. One GA–MLR model with six selected descriptors was obtained. This model could be used to predict the skin-irritant effect of the thietanes, with high statistical significance ($R^2_{\mathrm{training}} = 0.897$, $Q^2_{\mathrm{LOO}} = 0.872$, $Q^2_{\mathrm{LGO}} = 0.800$, $F = 87.253$, $R^2_{\mathrm{test}} = 0.921$). The results suggest that the number of bonds in the hydrogen-depleted molecule, electronegativity, mass, and neighbors of carbon atoms are the main independent factors contributing to the skin-irritant effect of the thietanes.

Keywords QSAR · Skin-irritant effect · Thietanes · Genetic algorithm · Stepwise

Introduction

Skin diseases and injuries are the most common job-related problems in industries such as manufacturing, food production, construction, machine tool operation, printing, metal plating, leather processing, engine service, landscaping, farming, and forestry. Because the skin is often exposed to chemicals, the potential for a chemical from a particular product to cause skin irritation must be evaluated as part of the overall safety-assessment process. The potential of chemicals or preparations to cause skin irritation has commonly been assessed by use of the rabbit Draize test [1, 2]. Although several experimental methods are available for screening the estrogenic activity of chemicals (e.g. in-vivo and in-vitro assay tests), these have all been carried out using receptors and other biological materials of human, rat, mouse, and calf origin [3]. They are costly, time-consuming, and can potentially produce toxic side products from the experimental methods used. One approach is to predict the skin-irritant effect on the basis of quantitative structure–activity relationships (QSARs) using structural data of the chemicals. QSAR has found diverse applications for predicting compounds' properties, including prediction of biological activity [4, 5], physical properties [6, 7], and toxicity [8, 9]. Several QSAR studies on skin-irritant effects of various compounds have been reported in recent years [10–13].

The thietanes are general synthetic products found in drugs, pesticides, and industrial chemicals to which skin may be exposed. Recent publications have shown that all the volatile compounds detected in the anal gland secretion of the Siberian weasel were sulfur-containing compounds (thietanes), for example 2,2-dimethylthietane, 2,4-dimethylthietane, 2,3-dimethylthietane, 2-ethylthietane, and 2-propylthietane [14], so sulfur-containing compounds such as thietanes can be used as malodorant components [15]. Assessment of the risk of skin irritation by these compounds is, therefore, highly significant. In this study, a quantitative structure–activity relationship was applied to

I. Kakoie Dinaki · M. Zarrineh (✉)
Department of Chemistry, Faculty of Science,
Imam Hossein University, Tehran, Iran
e-mail: zarrineh.m@gmail.com

the skin-irritant effects of a set of thietanes. Study of the structure–activity relationships of thietanes may provide knowledge enabling improvement of their properties.

The main objective of this work was to establish new QSAR models for predicting the skin-irritant effect of thietanes using the genetic algorithm–multiple linear regression technique (GA–MLR). The predicted skin-irritant effect was compared with that predicted by the stepwise multiple regression (SW-MLR) method and values obtained by use of PASS software.

Results and discussion

Data splitting

In order to build and test models, 84 compounds were separated into a training set of 67 compounds, which was used to build a model, and a test set of 17 compounds, which was used to test the built model. To assist separation of the data into training and test sets, the diversity of the sets was analyzed by principal-components analysis (PCA) of molecular descriptors of the compounds to detect homogeneities in the data sets and to show the spatial location of the samples. When PCA was used for analysis of the descriptors, PC1 and PC2 made contributions of 25.72 and 14.58%, respectively, to the total PCs. Figure 1 shows scatter distributions of the data for the first and second principal components. On inspection of this figure it was concluded that samples in both the training and test sets were evenly scattered in 2D space. This confirmed that it was feasible to split the data set. Moreover, the compounds in the training set were representative of all the data.

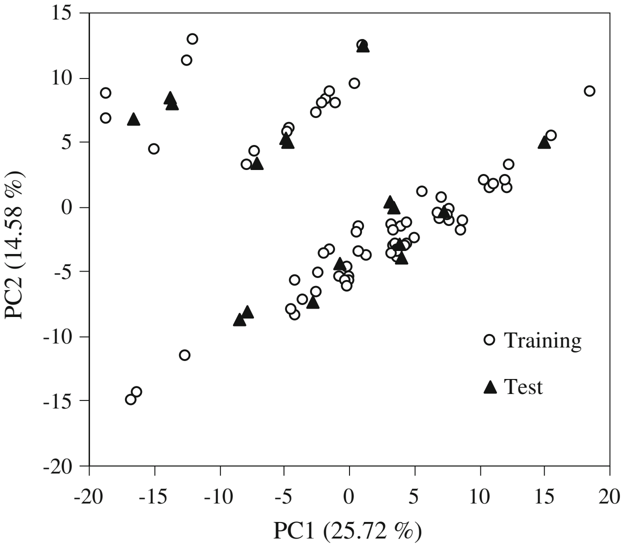

Fig. 1 Results from principal-components analysis of the training and the test sets

Variable selection and development of the models

We cannot have a-priori knowledge of which descriptors, and which particular combinations with others, are related to the studied response, so both stepwise multiple regression and the genetic algorithm were used for selection of the most important descriptors. First, MLR analysis with stepwise selection and elimination of variables was used to model the structure–activity relationships with a different set of descriptors.

This stepwise MLR analysis led to the derivation of one model, with six variables and good statistical data for the training set but with low generality and prediction ability for the test set. It is described by the equation:

$$\begin{aligned} Pa = {} & 94.115 - 3.791(\mathrm{Jhetm}) + 4.325(\mathrm{PJI2}) \\ & - 7.387(\mathrm{IC5}) + 61.962(\mathrm{R3m+}) - 131.636(\mathrm{R2v+}) \\ & - 22.910(\mathrm{R3e}) \end{aligned} \tag{1}$$

$N_{\mathrm{train}} = 67$, $R^2_{\mathrm{train}} = 0.881$, $F = 73.710$, $\mathrm{RMSE}_{\mathrm{train}} = 1.890$, $N_{\mathrm{test}} = 17$, $R^2_{\mathrm{test}} = 0.775$, $\mathrm{RMSE}_{\mathrm{test}} = 2.706$, $Q^2_{\mathrm{LOO}} = 0.848$, $Q^2_{\mathrm{LGO}} = 0.737$.

In this model the Jhetm descriptor is a Balaban-type index from the mass-weighted distance matrix, the PJI2 descriptor is a 2D Petitjean shape index, the IC5 descriptor is an information content index (neighborhood symmetry of 5-order), the R3m+ descriptor is R maximal autocorrelation of lag 3/weighted by atomic masses, the R2v+ descriptor is R maximal autocorrelation of lag 2/weighted by atomic van der Waals volumes, and the R3e descriptor is R autocorrelation of lag 3/weighted by atomic Sanderson electronegativities.

In this and the following equations, N is the number of compounds, R^2 is the squared correlation coefficient, F is the Fisher F statistic, RMSE is the root mean square error, and Q^2_{LOO} and Q^2_{LGO} are the squared cross-validation coefficients of leave-one-out (LOO) and leave-group-out (LGO).

The built model was used to predict the test set data. The prediction results are given in Table 1 and shown in Fig. 2. However, this procedure produced good results for the training set but did not produce good results for the test set. Therefore, the GA–MLR was used to select the best set of variables.

To select the optimum number of descriptors with the GA, the effect of the number of descriptors was investigated. The R^2 value can generally be increased by adding additional predictor variables to the model, even if the added variable does not contribute to reduction of the unexplained variance of the dependent variable. Therefore, use of R^2 requires special attention. For this reason, it is

Table 1 Calculated skin-irritant effect (*% Pa*) of 84 thietanes by PASS and the corresponding values predicted by the SW-MLR and GA–MLR methods

No.	Thietane	PASS	SW-MLR	GA–MLR
1	Thietane	23.6	22.5	23.3
2[a]	2-Methylthietane	12.5	13.3	13.4
3	3-Methylthietane	31.2	30.2	32.1
4	2-Ethylthietane	15.2	19.0	16.3
5	3-Ethylthietane	26.7	25.7	24.6
6	2-Propylthietane	11.1	11.8	12.8
7	3-Propylthietane	19.6	20.9	19.1
8	2-Phenylthietane	16.7	16.3	15.2
9	3-Phenylthietane	23.2	21.3	21.6
10	2-Benzylthietane	12.2	13.1	14.2
11	3-Benzylthietane	18.5	19.3	19.9
12	2-Ethyl-2-methylthietane	20.8	20.1	19.5
13	2,2-Dimethylthietane	17.6	16.6	20.0
14	2,3-Dimethylthietane	17.0	18.5	20.3
15	2,4-Dimethylthietane	15.1	15.2	13.7
16	3,3-Dimethylthietane	36.9	32.8	34.1
17	2-Ethyl-3-methylthietane	23.2	21.2	20.0
18	2-Ethyl-4-methylthietane	14.0	11.7	11.6
19[a]	3-Ethyl-3-methylthietane	32.6	28.3	29.5
20	2,3-Diethylthietane	17.7	17.6	17.7
21	2,2-Diethylthietane	23.5	23.9	24.1
22	3,3-Diethylthietane	31.7	29.5	31.1
23	2,4-Diethylthietane	18.5	18.1	18.0
24	3-Ethyl-2,2-dimethylthietane	20.8	19.4	19.5
25	4-Ethyl-2,2-dimethylthietane	14.0	13.3	14.0
26	2-Ethyl-3,3-dimethylthietane	26.4	23.7	23.4
27	2-Ethyl-3,4-dimethylthietane	19.3	15.2	16.0
28[a]	3-Ethyl-2,4-dimethylthietane	16.6	15.8	17.1
29	2-Hexylthietane	8.7	7.9	7.2
30	3-Hexylthietane	14.8	16.9	13.5
31	2,3-Diethyl-4-methylthietane	14.4	14.7	12.9
32	2,2-Dibenzylthietane	15.9	12.8	15.8
33	3,3-Dibenzylthietane	22.1	18.8	20.4
34	2,3-Dibenzylthietane	12.6	13.4	13.7
35[a]	2,4-Dibenzylthietane	13.9	11.1	10.4
36	3,3-Diethyl-2-methylthietane	21.7	23.2	24.2
37	2,2-Diethyl-3-methylthietane	25.9	23.6	25.7
38[a]	2-Butylthietane	9.8	11.1	10.1
39[a]	3-Butylthietane	17.0	21.4	16.5
40	2-Pentylthietane	8.7	7.9	8.7
41	3-Pentylthietane	14.8	17.4	15.0
42	2,3-Dipropylthietane	13.0	13.0	13.2
43	3,3-Dipropylthietane	23.6	26.2	23.9
44	2,2-Dipropylthietane	16.7	19.1	18.3
45[a]	2,2-Dibutylthietane	14.5	16.2	13.8
46	2,4-Dipropylthietane	12.5	12.7	12.8
47[a]	3,3-Dibutylthietane	20.4	22.5	20.5
48[a]	3-Ethyl-2-phenylthietane	18.3	14.6	17.0
49	2-Ethyl-2-propylthietane	15.9	15.5	16.0
50[a]	3-Ethyl-3-propylthietane	21.8	22.8	23.2
51	2-Methyl-3-propylthietane	12.2	13.6	15.2
52	2-Methyl-2-propylthietane	15.2	14.9	15.0
53	2-Methyl-4-propylthietane	10.0	9.7	9.0
54	3-Methyl-2-propylthietane	14.3	17.7	17.0
55	2-Ethyl-4-propylthietane	12.2	8.6	10.5
56	3-Ethyl-2-methylthietane	16.1	16.2	17.3
57	2-Ethyl-3-propylthietane	14.3	16.8	17.7
58	3-Ethyl-2-propylthietane	14.3	16.0	16.4
59	2,2,3-Trimethylthietane	23.4	22.8	23.0
60	2,3,3-Trimethylthietane	23.9	24.0	25.7
61	2,2,4-Trimethylthietane	10.1	13.4	14.5
62	3-Ethyl-2-phenylthietane	18.3	16.7	18.0
63	2,2-Diethyl-4-methylthietane	16.3	19.7	16.9
64	2,3,4-Triethylthietane	18.7	19.7	15.5
65	2,2,3-Triethylthietane	20.4	23.9	21.8
66[a]	2,2,4-Triethylthietane	15.5	17.4	17.4
67	2,3,3-Triethylthietane	21.2	24.0	21.9
68	2-Phenyl-3-propylthietane	14.6	13.6	13.9
69	3-Phenyl-2-propylthietane	15.3	16.2	15.1
70[a]	3-Ethyl-2,3-dimethylthietane	23.5	22.8	21.9
71	2,4-Diphenylthietane	14.8	16.5	16.3
72[a]	2,2-Diphenylthietane	19.2	23.4	19.4
73	2,3-Diphenylthietane	13.0	13.6	16.3
74[a]	3,3-Diphenylthietane	26.0	23.0	24.6
75	2-Methyl-4-phenylthietane	12.9	13.6	12.9
76	2-Methyl-3-phenylthietane	16.7	14.5	15.7
77	3-Methyl-2-phenylthietane	20.0	19.8	18.1
78	2-Ethyl-3-phenylthietane	19.4	18.5	16.8
79[a]	2-Ethyl-4-phenylthietane	15.0	13.2	13.0
80[a]	2-Phenyl-4-propylthietane	11.6	9.0	11.3
81	2,2-Dihexylthietane	12.6	11.2	12.0
82	3,3-Dihexylthietane	17.6	16.5	17.1
83[a]	2,2-Dipentylthietane	12.6	16.3	15.2
84	3,3-Dipentylthietane	17.6	19.4	18.1

[a] Used as test set

better to use another statistical variable, named the adjusted R^2 (R^2_{adj}).

R^2_{adj} is defined as:

$$R^2_{\mathrm{adj}} = 1 - \left(1 - R^2\right)\left(\frac{n-1}{n-p-1}\right) \qquad (2)$$

In this equation, n is the number of compounds and p is the number of variables. R^2_{adj} is interpreted similarly to the R^2 value, considering the number of degrees of freedom also. It is adjusted by dividing the residual sum of squares

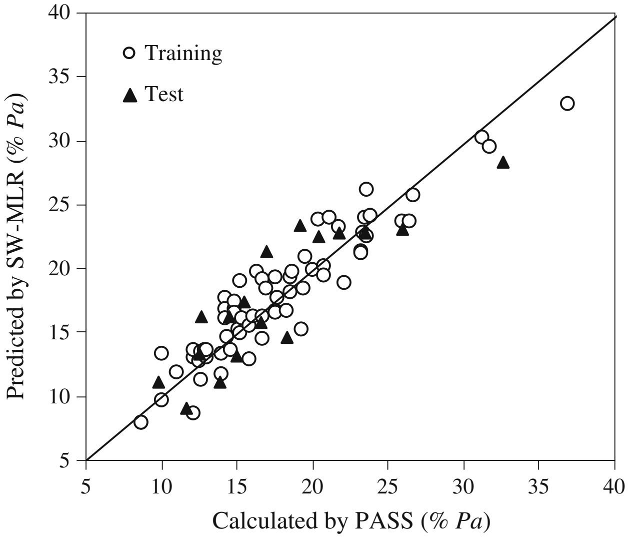

Fig. 2 Calculated versus predicted (by SW-MLR) skin-irritant effect (% *Pa*) of the thietanes

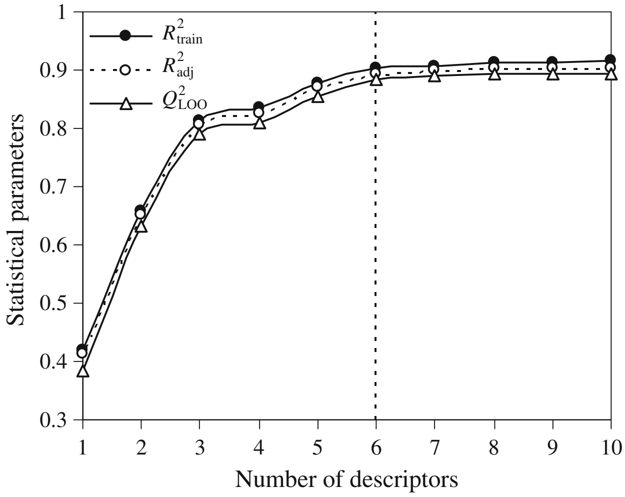

Fig. 3 Effect of the number of descriptors on: R^2_{train}, R^2_{adj}, and Q^2_{LOO} of the GA–MLR model

and total sum of squares by their respective degrees of freedom. The R^2_{adj} value diminishes if addition of a variable to the equation does not reduce the unexplained variance [16]. Subsequently, R^2_{adj} is used to compare models with different numbers of predictor variables. Another statistical variable is the LOO cross-validation correlation coefficient (Q^2_{LOO}). Figure 3 shows plots of R^2, R^2_{adj}, and Q^2_{LOO} for the training set as a function of the number of descriptors for the 1–10-descriptor models. R^2, R^2_{adj}, and Q^2_{LOO} increase with increasing number of descriptors.

Because it can be seen that models with 7–10 descriptors did not significantly improve the statistics of the model, it was determined that the optimum subset size had been achieved with a maximum of six descriptors. With the six descriptors selected by use of the GA, we built a linear model using the training set, and the following equation was obtained:

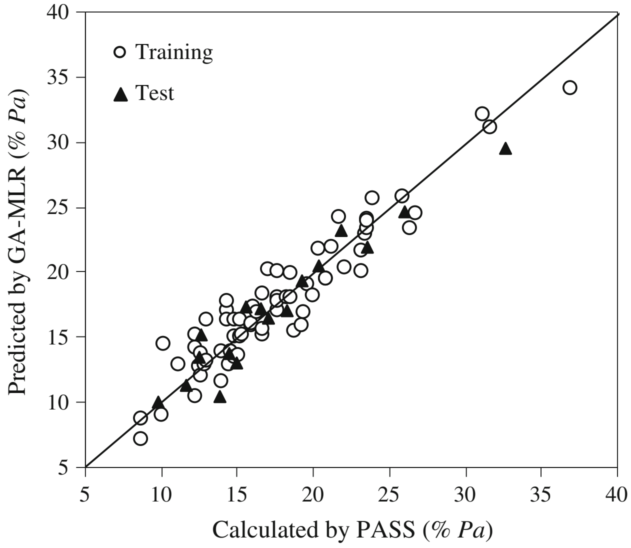

Fig. 4 Calculated versus predicted (by GA–MLR) skin-irritant effect (% *Pa*) of thietanes

$$\begin{aligned} Pa = {} & 72.204 - 158.920(\mathrm{X2A}) - 5.584(\mathrm{IC5}) \\ & - 5.310(\mathrm{Mor18e}) + 8.395(\mathrm{G2m}) \\ & + 8.300(\mathrm{HATS4u}) + 2.932(\mathrm{C\text{-}006}) \end{aligned} \tag{3}$$

$N_{train} = 67, R^2_{train} = 0.897, F = 87.253, \mathrm{RMSE}_{train} = 1.753, N_{test} = 17, \quad R^2_{test} = 0.921, \quad \mathrm{RMSE}_{test} = 1.650, \quad Q^2_{LOO} = 0.872, Q^2_{LGO} = 0.799.$

The built model was used to predict the test set data. The prediction results are given in Table 1. As can be seen from this table, the predicted values of *Pa* are in good agreement with the PASS-obtained values. The predicted values of *Pa* for the compounds in the training and test sets using Eq. (3) have been plotted versus the PASS-obtained values in Fig. 4.

Validation and comparison of the models

The applicability domain of these models was evaluated by leverage analysis expressed as Williams plots (Figs. 5, 6) in which the standardized residuals and the leverage values (*h*) were plotted. From Figs. 5 and 6 it is obvious that there are only three chemicals in the SW-MLR model (**2**, **36**, **44**) and three chemicals in the GA–MLR model (**1**, **3**, **16**) that have higher leverage than the warning h^* value of 0.313, thus they can be regarded as structural outliers. Fortunately, in these cases the values predicted by the models are good for these compounds, thus they are ‘‘good leverage’’ chemicals. As can be seen from these figures there is no outlier compound with standard residuals >3δ.

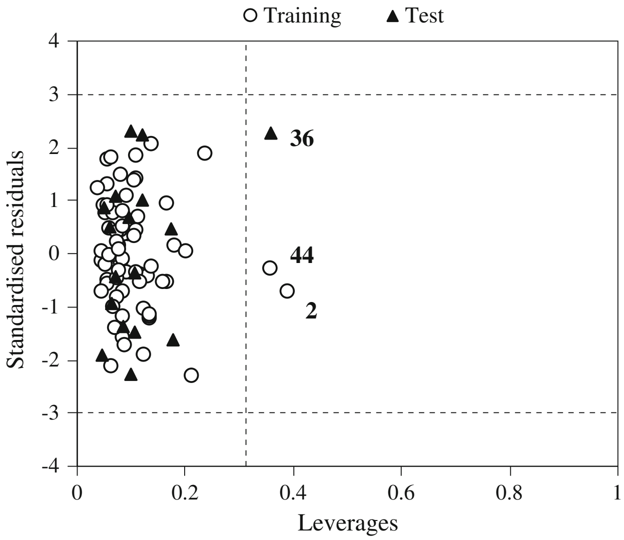

Fig. 5 Williams plot of SW-MLR model. The training and test set samples are labeled differently. The *dashed lines* are the 3δ limit and the warning value of hat ($h^* = 0.313$)

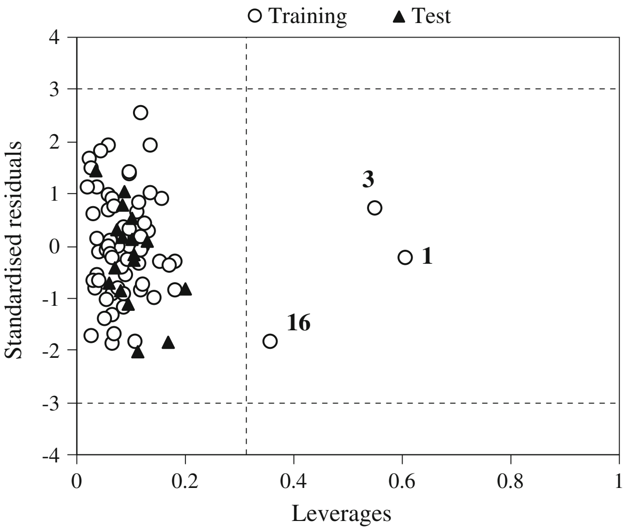

Fig. 6 Williams plot of GA–MLR model. The training and test set samples are labeled differently. The *dashed lines* are the 3δ limit and the warning value of hat ($h^* = 0.358$)

The models obtained were also validated by use of the LOO and LGO cross-validation processes. For LOO cross-validation, a data point is removed from the set, and the model is recalculated. The predicted activity for that point is then compared with its PASS-obtained value. This is repeated until each data point has been omitted once. For LGO, 20% of the data points are removed from the datasets and the models refitted; the predicted values for those points are then compared with its PASS values. Again, this is repeated until each data point has been omitted once. Cross-validation data for two models are shown below Eqs. (1) and (3). These indicate that the GA–MLR model obtained has good internal and external predictive power, but the SW-MLR model only has a good internal predictive power.

Table 2 R^2_{train} and Q^2_{LOO} values of the GA–MLR model after several Y-randomization tests

Iteration	R^2_{train}	Q^2_{LOO}
1	0.084	0.001
2	0.050	0.054
3	0.006	0.214
4	0.091	0.005
5	0.135	0.009
6	0.118	0.005
7	0.066	0.017
8	0.063	0.017
9	0.004	0.429
10	0.099	0.002

Statistical data for the results obtained from the two studied models for the same set of compounds are presented below Eqs. (1) and (3). It is apparent that RMSE of the GA–MLR model for the training and test data sets were lower than those of the model proposed in the SW-MLR method. The squared correlation coefficient R^2 given by GA–MLR was much higher than of SW-MLR in terms of the training and test sets. The results of the F-test were obtained and are also shown with these equations. It can be seen that the GA–MLR model yields higher F values, so this model gives more satisfactory results than the SW-MLR method. Consequently, this GA–MLR approach is currently the most accurate method for prediction of the skin-irritant effect of these thietanes.

Because the statistical data obtained indicated the superiority of the genetic algorithm over the stepwise multiple regression method in feature selection, for more validation only the GA–MLR model was investigated.

In order to assess the robustness of the GA–MLR model, the Y-randomization test was applied [17]. The dependent variable vector (*Pa*) was randomly shuffled and a new QSAR model developed using the original independent variable matrix. The new QSAR models (after several repetitions) were expected to have low R^2 and Q^2_{LOO} values (Table 2). If the opposite happens then an acceptable QSAR model cannot be obtained for the specific modeling method and data. As can be seen from Table 2, the model was reasonable, and had not been obtained by chance.

The correlation matrix of these descriptors is shown in Table 3. It is apparent that the value of the correlation coefficient for each pair of descriptors was less than 0.75, which meant that the selected descriptors were independent.

Table 3 Correlation coefficient matrix of the selected descriptors obtained by GA–MLR

	X2A	IC5	Mor18e	G2m	HATS4u	C-006
X2A	1					
IC5	−0.441	1				
Mor18e	0.408	−0.744	1			
G2m	0.422	−0.450	0.252	1		
HATS4u	0.341	−0.636	0.557	0.234	1	
C-006	0.145	−0.095	−0.146	0.303	0.145	1

Description of descriptors

Besides being statistically significant, QSAR models should also provide useful chemical insight into the mechanism of activity. For this reason, an acceptable interpretation of QSAR results is provided below. By interpreting the descriptors contained in the model, it is possible to gain insight into which factors are related to the skin-irritant effect of thietanes. X2A is one of the topological descriptors which appeared in the GA–MLR model. X2A is the average connectivity index, chi-2, that is calculated from the hydrogen-depleted molecular graph. The average connectivity index depends on the number of bonds in the hydrogen-depleted molecule. As can be seen from Eq. (3), the coefficient of X2A has a negative sign, which indicates that the skin-irritant effect is inversely related to this descriptor. The second descriptor of the model, IC5, is one of the topological descriptors and constitutes the information content index (neighborhood symmetry of 5-order). The information content indices [18] are calculated on the basis of the pairwise equivalence atoms in a hydrogen-filled molecule. The negative sign of IC5 indicates that the skin-irritant effect is inversely related to this descriptor. Mor18e is the third descriptor appearing in the model. It is one of the 3D MoRSE descriptors. 3D MoRSE descriptors (3D molecule representation of structures based on electron diffraction) are derived from IR spectra simulation using a generalized scattering function [18]. This descriptor was proposed as signal 18/weighted by atomic Sanderson electronegativities which relates to the electronegativity of the molecules. Mor18e has a negative sign, which indicates that the skin-irritant effect is inversely related to this descriptor. G2m (2nd component accessibility directional WHIM index/weighted by atomic masses) is the fourth descriptor appearing in the model. It is one of the WHIM descriptors which are based on the statistical indices calculated from projections of atoms along principal axes. The algorithm consists of performing principal-components analysis on the centered cartesian coordinates of a molecule by using a weighted covariance matrix obtained from different weighing schemes for the atoms. Directional WHIM symmetry descriptors are related to the number of central symmetric atoms (along the *m*th component), the number of asymmetric atoms, and the total number of atoms of the molecule. The atomic masses are one of the weighting schemes that are used for computing the weighted covariance matrix in this descriptor (G2m). The G2m has a positive sign which indicates that skin-irritant effect is directly related to this descriptor; therefore, increasing the molecular mass leads to increased skin-irritative effect. HATS4u belongs to the GETAWAY descriptors and represents leverage-weighted autocorrelation of lag 4/unweighted. HATS4u has a positive sign in the model. C-006 (CH2RX, atom-centered fragments) is another descriptor appearing in model. It is one of the atom-centered fragment descriptors that describe each atom by its own atom type and the bond types and atom types of its first neighbors. This descriptor is defined by looking at the first neighbors of carbon atoms. The neighbors of a carbon atom in this case can be hydrogen (represented as H), carbon (represented as R), or hetero atoms (represented as X). C-006 has a positive sign, which indicates that skin-irritant effect is directly related to this descriptor.

Conclusion

QSAR analysis was performed on a series of thietanes using the MLR procedure. The best set of calculated descriptors was selected with the stepwise and genetic algorithms. The two methods had good statistical quality for the training set. GA–MLR was found to be superior to stepwise MLR with reference to the external set predictions. The number of bonds in the hydrogen-depleted molecule, electronegativity, molecular mass, and neighbors of carbon atoms were found to be important factors affecting the skin-irritant effect of the thietanes.

Methods

Data set

Skin-irritant effect was predicted with PASS [19, 20]. PASS version 1.603 predicts 900 types of biological activity on the basis of the structural formulae of compounds. Prediction is based on a structure–activity relationship knowledge base developed by analysis of a training set containing more than 45,000 known biologically active compounds. The result of prediction is presented as a list of activity with appropriate *Pa*, which is the estimate of the probability of being active. Eighty-four thietane derivatives are considered in this study and their

Pa values are used for subsequent QSAR analyses as the response variables. Chemical names and activity data for the complete set of thietanes (divided into the corresponding training and test sets based on principal-components analysis) are presented in Table 1.

Calculation of the descriptors

The first step in obtaining a QSAR model was to encode the structural features of the molecules, which were named molecular descriptors. The molecular descriptors used to search the best model for the skin-irritant effect of these compounds were calculated by use of Dragon software [21] on the basis of minimum-energy molecular geometries. These geometries were optimized with the aid of the HyperChem package [22], based on the AM1 semiempirical method. The calculated descriptors were first analyzed for the existence of constant or near-constant variables. Those detected were then removed. In addition, to reduce the redundancy existing in the descriptor data matrix, the descriptors' correlations with each other and with the *Pa* of the molecules were examined. Afterwards collinear descriptors (i.e. $r > 0.9$) were detected. Among the collinear descriptors the one with the highest correlation with *Pa* was retained. The other descriptors were removed from the data matrix. Remaining descriptors were considered for further investigations after discarding the descriptors with constant and intercorrelated ones.

Genetic algorithm (GA)

Nowadays the GA is well-known as an interesting, and the most widely used, variable-selection method. The GA is a stochastic method for solving optimization problems defined by fitness criteria, by applying the evolution hypothesis of Darwin and different genetic functions, i.e. crossover and mutation. To select the most relevant descriptors, the evolution of the population was simulated [23–25]. The population of the first generation was selected randomly. Each individual member of the population, defined by a chromosome of binary values, is represented a subset of descriptors. The number of the genes on each chromosome was equal to the number of the descriptors. A gene was given the value of 1 if its corresponding descriptor was included in the subset; otherwise, it was given the value of zero. The number of genes with the value 1 was kept relatively low to have a small subset of descriptors [26]. As a result, the probability of generating 0 for a gene was set greater (at least 60%) than for the value of 1. The operators used here were crossover and mutation. The application probability of these operators was varied linearly with a generation renewal (0–0.1% for mutation and 60–90% for crossover). The population size was varied between 50 and 250 for different GA runs. For a typical run, evolution of the generation was stopped when 90% of the generations took the same fitness. The GA program was written in Matlab 6.5 [27].

Applicability domain

The chemical domain of the chemicals studied by use of the models was verified by the leverage approach to verify prediction reliability [17, 28]. No matter how robust, significant, and validated a QSAR model may be, it cannot be expected to reliably predict the modeled activity for all chemicals. Therefore, a defined applicability domain (AD) is absolutely necessary before a model is put into use for screening chemicals, and predictions only for chemicals that fall in this domain may be regarded as reliable [17, 28]. This problem is commonly solved by use of leverage (h) [29]. The warning leverage h^* is fixed at $3k'/n$ generally, where n is the number of training compounds and k' is the number of model descriptors plus one. A leverage greater than h^* means that the predicted response is the result of substantial extrapolation of the model and may not be reliable. Moreover, the plot of leverages (hat diagonals) versus standardized residuals, i.e. the Williams plot, can verify the presence of response outliers (i.e. compounds with cross-validated standardized residuals greater than three standard deviation units, $\pm 3\delta$) and chemicals very structurally influential in determining model parameters.

References

1. Draize JH, Woodard G, Calvery HO (1944) J Pharmacol Exp Ther 82:77
2. Draize JH (1959) Dermal toxicity. Appraisal of the safety of chemicals in foods, drugs, and cosmetics. Association of foods and drugs officials of the United States. Littleton, CO, pp 46–59
3. Hill DL (1972) The biochemistry and physiology of Tetrahymena, 1st edn. Academic Press, New York
4. Riahi S, Pourbasheer E, Dinarvand R, Ganjali MR, Norouzi P (2008) Chem Biol Drug Des 72:575
5. Riahi S, Pourbasheer E, Ganjali MR, Norouzi P (2008) Chem Biol Drug Des 72:205
6. Riahi S, Ganjali MR, Pourbasheer E, Norouzi P (2008) Chromatographia 6:917
7. Riahi S, Pourbasheer E, Ganjali MR, Norouzi P (2009) J Hazard Mater 166:853
8. Khadikar PV, Phadnis A, Shrivastava A (2002) Bioorg Med Chem 10:1181
9. Agrawal VK, Khadikar PV (2001) Bioorg Med Chem 9:3035
10. Bratt MD (1996) Toxicol In Vitro 10:247
11. Golla S, Madihally S, Robinson RL Jr, Gasem KAM (2009) Toxicol In Vitro 23:176
12. Hayashi M, Nakamura Y, Higashi K, Kato H, Kishida F, Kaneko H (1999) Toxicol In Vitro 13:915
13. Kodithala K, Hopfinger AJ, Thompson ED, Robinson MK (2002) Toxicol Sci 66:336
14. Zhang JX, Sun LX, Zhang ZB, Wang ZW, Chen Y, Wang R (2002) J Chem Ecol 28:1287

15. Shih M, David LL, Lampi KJ, Ma H, Fukiage C, Azuma M, Shearer TR (2001) Curr Eye Res 22:458
16. Hansch C, Taylor J, Sammes P (1990) Comprehensive medicinal chemistry: the rational design, mechanistic study and therapeutic application of chemical compounds. vol 6. Pergamon, New York, pp 1–19
17. Tropsha A, Gramatica P, Gombar VK (2003) QSAR Comb Sci 22:69
18. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley, Weinheim
19. Raymond SA, Steffensen SC, Gudino LD, Strichartz GR (1989) Anesth Analg 68:563
20. Stepanchikova AV, Lagunin AA, Filimonov DA, Poroikov VV (2003) Current Med Chem 10:225
21. Todeschini R, Consonni V, Pavana M (2002) Dragon, Software version 2.1. http://www.disat.unimib.it/chm/
22. HyperChem Release 7. HyperCube, Inc, Alberta, Canada; http://www.hyper.com
23. Hunger J, Huttner G (1999) J Comput Chem 20:455
24. Ahmad S, Gromiha MM (2003) J Comput Chem 24:1313
25. Waller CL, Bradley MP (1999) J Chem Inf Comput Sci 39:345
26. Aires-de-Sousa J, Hemmer MC, Gasteiger J (2002) Anal Chem 74:80
27. The Mathworks Inc (2002) Genetic algorithm and direct search toolbox users guide, Massachusetts
28. Gramatica P (2007) QSAR Comb Sci 26:694
29. Atkinson AC (1985) Plots, transformations and regression: an introduction to graphical methods of diagnostic regression analysis. Clarendon Press, Oxford