Available on line www.jocpr.com

Journal of Chemical and Pharmaceutical Research

ISSN No: 0975-7384
CODEN(USA): JCPRC5

J. Chem. Pharm. Res., 2011, 3(1):93-102

A simple and readily integratable approach to toxicity prediction for some of the anti cancer drugs

Z. Bayat* and S. Vahdani

Department of Chemistry, Islamic Azad University-Quchan Branch, Iran

ABSTRACT

A quantitative structure–activity relationship (QSAR) study was performed to develop models those relate the structures of 13anti-cancer drugs to inhibit 50% of sensitive cell growth (pLD50). The aim of this paper is to establish a predictive model for pLD50 penetration using simple descriptors. The usefulness of the quantum chemical descriptors, calculated at the level of the DFT theories using 6-31G basis set for QSAR study of anti-cancer drugs was examined. The QSAR model developed contributed to a mechanistic understanding of the investigated biological effects. Multiple Linear Regressions (MLR) was employed to model the relationships between molecular descriptors and biological activities of molecules using stepwise method and genetic algorithm as variable selection tools. A multi-parametric equation containing maximum two descriptors at B3LYP/6-31G* method with good statistical qualities (R^2_{train}=0.932, F_{train}=69.225, Q^2_{LOO}=0.889,R^2_{adj}=0.919,Q^2_{LGO}=0.893) was obtained by Multiple Linear Regression using stepwise method. The accuracy of the proposed MLR model was illustrated using the following evaluation techniques: cross-validation, and Y-randomisation. The predictive ability of the model was found to be satisfactory and could be used for designing a similar group of* .cancer drugs-anti

keywords: Anti-cancer drugs,QSAR,DFT,LD50,MLR.

INTRODUCTION

Anthracycline antibiotics such as doxorubicin and its analogues have been in common use as anticancer drugs for almost half a century. There has been intense interest in the DNA binding sequence specificity of these compounds in recent years, with the hope that a compound could be identified that could possibly modulate gene expression or exhibit reduced toxicity. The experimental measurement of the inhibition activity of chemicals is difficult, expensive and time-consuming, thus a great deal of effort has been put into attempting the estimation of activity through statistical modeling. Quantitative structure–activity relationship (QSAR) analysis is an effective method in research into rational drug design and the mechanism of drug actions. In addition, it is useful in areas like the design of virtual compound libraries and the computational-chemical optimisation of compounds. QSAR studies can express the biological activities of compounds as a function of their various structural parameters and also describes how the variation in biological activity depends on changes in the chemical structure [1]. Recently, a QSAR study of biological activity has been published by our group [2–4]. If such a relationship can be derived from the structure-activity data, the model equation allows medicinal chemists to say with some confidence which properties are important in the mechanism of drug action. The success of a QSAR study depends on choosing robust statistical methods for producing the predictive model and also the relevant structural parameters for expressing the essential features within those chemical structures. Nowadays, genetic algorithms (GA) are well known as interesting and widely used methods for variable selection [5-11]. GA are stochastic methods used to solve the optimisation problems defined by the fitness criteria, applying the evolutionary hypothesis of Darwin and also different genetic functions i.e. crossover and mutation. In the present work, we have used a genetic algorithm for the variable selection, and developed an MLR model for the QSAR analysis of the cancer-anti drugs.

In a QSAR study the model must be validated for its predictive value before it can be used to predict the response of additional chemicals. Validating QSAR with external data (i.e. data not used in the model development), although demanding, is the best method for validation. In the present work, the data splitting was performed randomly and was confirmed by the factor spaces of the descriptors, as in our previous work [13, 14]. Finally, the accuracy of the proposed model was illustrated using the following: leave one out,cross-validations and Y-randomisation techniques.

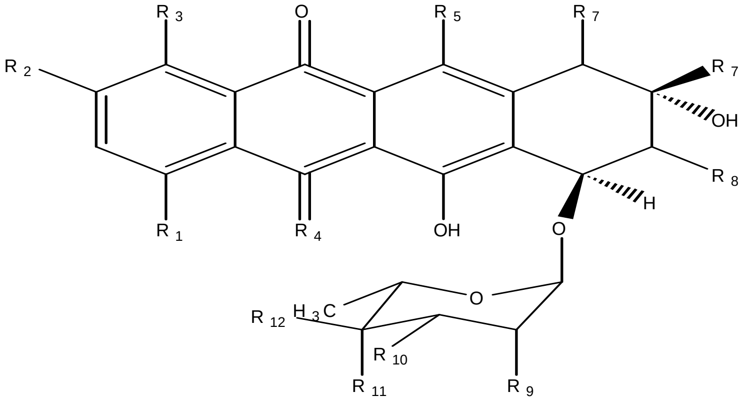
EXPERIMENTAL SECTION

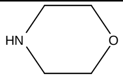
Methodology

Data set

In this study, the data set of 13 anti-cancer drugs [14, 15]. The inhibitory activity values are expressed as the inhibit 50% of sensitive cell growth (pLD50).The chemical structures and activity data for the complete set of compounds are presented in Table 1. The activity data [pLD50 (μM)] was converted to the logarithmic scale pLD50 [-log LD50 (M)] and then used for the subsequent QSAR analyses as the response variables.

Table 1. Chemical structures and the corresponding observed and predicted pLD50 values by the MLR method.



N	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}	R_{11}	R_{12}	Exp.	Pread	Red
1	OCH_3	H	H	O	OH	H	$COCH_2OH$	H	H	NH_2	OH_{ndo}	H	21.8	20.9	15
2	OCH_3	H	H	O	OH	H	$COCH_3$	H	H	NH_2	OH	H	20	20.9	15
3	H	H	H	O	OH	H	$COCH_3$	H	H	OH	OH	H	16.2	.16,1	15
4	OCH_3	H	H	O	OH	H	$COCH_2OCH_3$	H	H	NH_2	OH	H	14.2	15.6	15
5	OCH_3	H	H	O	OH	H	$COCH_2OH$	H	H	NH_2	H	H	14.1	15.8	15
6	OCH_3	H	H	O	OH	H	$COCH_3$	H	H	NH_2	H	H	17.9	15.4	15
7	OCH_3	H	H	O	OH	H	$COCH_2OH$	H	H		OH	H	17	15.7	14
8	OH	H	H	O	OH	$COOCH_3$	CH_2CH_3	H	H	$N(CH_3)_2$	OH	H	36	36.1	14
9	OCH_3	H	H	O	OH	H	$COCH_2OCO(CH_2)_3CH_3$	H	H	$NHCOCF_3$	OH	H	13.9	15.4	14
10	OH	H	H	O	OH	H	$COCH_3$	H	H	NH_2	H	H	13.5	16	14
11	OH	H	HO	O	H	H	$COCH_2OH$	H	H	NH_2	N(CH3)2	H	18.7	18.1	15
12	H	H	OH	O	H	CO_2CH_3	CH_2CH_3	H	H	$N(CH_3)_2$	OH	H	16.5	15.7	14
13	CH_3	H	H	O	OH	H	$COCH_3$	H	H	NH_2	H	H	18.7	16.6	14

a= mg/kg

Table2. The calculated descriptors used in this study.

Abbreviation	Symbol	Descriptors	Abbreviation	Symbol	Descriptors
E GAP	difference between LUMO and HOMO	Quantum chemical descriptors	MDP	Molecular Dipole Moment	Quantum chemical descriptors
H	Hardness [η=1/2 (HOMO+LUMO)]		MP	Molecular Polarizability	
S	Softness (S=1/ η)		NPA	Natural Population Analysis	
X	Electro negativity [χ= -1/2 (HOMO–LUMO)]		EP	Electrostatic Potentialc	
Ω	El Electro philicity (ω=χ2 /2 η)		HOMO	Highest Occupied Molecular Orbital	
MC	Mullikenl Chargeg		LUMO	Lowest Unoccupied Molecular Orbital	
SA	Molecule surface area	Chemical properties	Log P	Partition Coefficient	Chemical properties
HE	Hydration Energy		M	Mass	
REF	Refractivity		V	Molecule volume	

Molecular descriptor generation

All of the molecules were drawn into the Hyper Chem. The Gaussian 03 package was used for calculating the molecular descriptors. Some of the descriptors are obtained directly from the chemical structure, e. g. constitutional, geometrical, and topological descriptors. Other chemical and physicochemical properties were determined by the chemical structure (lipophilicity, hydrophilicity descriptors, electronic descriptors, energies of interaction). In this work, we used Gaussian 03 for abinitio calculations. DFT method at 6-31G* were applied for optimization of anti-cancer drugs and calculation of many of the descriptors. A large number of descriptors were calculated by Gaussian package and Hyperchem software (Table2).One way to avoid data redundancy is to exclude descriptors that are highly intercorrelated with each other before performing statistical analysis.

Genetic algorithm

Genetic algorithms (GAs) are governed by biological evolution rules [16]. These are stochastic optimization methods that have been inspired by evolutionary principles. The distinctive aspect of a GA is that it investigates many possible solutions simultaneously, each of which explores a different region in the parameter of space [17]. To select the most relevant descriptors, the evolution of the population was simulated [18-20]. The first generation population was randomly selected; each individual member in the population was defined by a chromosome of binary values and represented a subset of descriptors. The number of the genes at 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 the genes with the value of 1 was kept relatively low to have a small subset of descriptors [21]. As a result, the probability of generating 0 for a gene was set greater (at least 60 %) than the value of 1. The operators used here were the crossover and mutation operators. 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 the different GA runs. For a typical run, the evolution of the generation was stopped when 90% of the generations took the same fitness. The fitness function used here was the leave-one-out cross validated correlation coefficient, $Q2_{LOO}$. The GA program was written in Matlab 6.5 [22].

RESULTS AND DISCUSSION

The diversity of the training set the was analysed using the principal component analysis (PCA) method. The PCA was performed with the calculated structure descriptors for the whole data set to detect the homogeneities in the data set. The PCA results showed that three principal components (PC1and PC2) described 73.05% of the overall variables, as follows: PC1 = 54.64% and PC2 = 45.35%. Since almost all the variables can be accounted for by the first three PCs, their score plot is a reliable representation of the spatial distribution of the points for the data set. The plot of PC1and PC2 displayed the distribution of compounds over the first three principal components space.

The multi-collinearity between the above seven descriptors were detected by calculating their variation inflation factors (VIF), which can be calculated as follows:

$$\mathrm{VIF}=\frac{1}{1-r2} \qquad (1)$$

Where r is the correlation coefficient of the multiple regression between the variables in the model. If VIF equals to 1, then no inter-correlation exists for each variable; if VIF falls into the range of 1–5, the related model is acceptable; and if VIF is larger than 10, the related model is unstable and a recheck is necessary [23]. The corresponding VIF values of the seven descriptors are shown in (Table 3). As can be seen from this table, most of the variables had VIF values of less than 5, indicating hat the obtained model has statistic significance. To examine the relative importance as well as the contribution of each descriptor in the model, the value of the mean effect (MF) was calculated for each descriptor. This calculation was performed with the equation below:

$$\mathrm{MF}_j=\frac{\beta \sum_{i=1}^{i=n} dij}{\sum_j^m \beta j \sum_i^n \beta ij} \qquad (2)$$

Where *MFj* represents the mean effect for the considered descriptor *j, βj* is the coefficient of the descriptor *j, dij* stands for the value of the target descriptors for each molecule and, eventually, *m* is the descriptors number for the model. The MF value indicates the relative importance of a descriptor, compared with the other descriptors in the model. Its sign indicates the variation direction in the values of the activities as a result of the increase (or reduction) of the descriptor values. The mean effect values are shown in Table 3.

Table 3. The linear model based on the tree parameters selected by the GA-MLR method.

Descriptor	Chemical meaning	MFa	VIFb
Constant	Intercept	0	0
MC1	Mullliken charge 1	12.61145	1.008694
MC24	Mulliken charge 24	-11.6114	1.008694

[a] *Mean effect*

[b] *Variation inflation factors*

In a QSAR study, generally, the quality of a model is expressed by its fitting ability and prediction ability, and of these the prediction ability is the more important. With the selected descriptors, we have built a linear model using the set data, and the following equation was obtained:

$$\mathrm{pLD50}=18.79(\pm 1.399) +41.62\mathrm{MC}_1 (\pm 3.66) + 34.12\mathrm{MC}_{24} (\pm 8.36) \qquad (3)$$

$$N=13 \quad R^2=0.933 \; F=69.224 \quad R^2_{adj}= 0.919 \quad Q^2_{LOO}=0.899 \quad Q^2_{LGO}=0.893$$

In this equation, N is the number of compounds, R^2 is the squared correlation coefficient, $Q^2{}_{LOO}$, $Q^2{}_{GLO}$ are the squared cross-validation coefficients for leave one out, bootstrapping and external test set respectively, F is the Fisher F statistic. The figures in parentheses are the standard deviations. As can be seen from Table 1, the calculated values for the pLD50 are in good agreement with those of the experimental values figure1. A plot of the residual for the predicted values of pLD50 for both the training and test sets against the experimental pLD50 values are shown in Figure 2. As can be seen the model did not show any proportional and systematic error, because the propagation of the residuals on both sides of zero are random. The real usefulness of QSAR models is not just their ability to reproduce known data, verified by their fitting power (R^2), but is mainly their potential for predictive application. For this reason the model calculations were performed by maximising the explained variance in prediction, verified by the leave-one-out cross-validated correlation coefficient,Q^2_{LOO}.

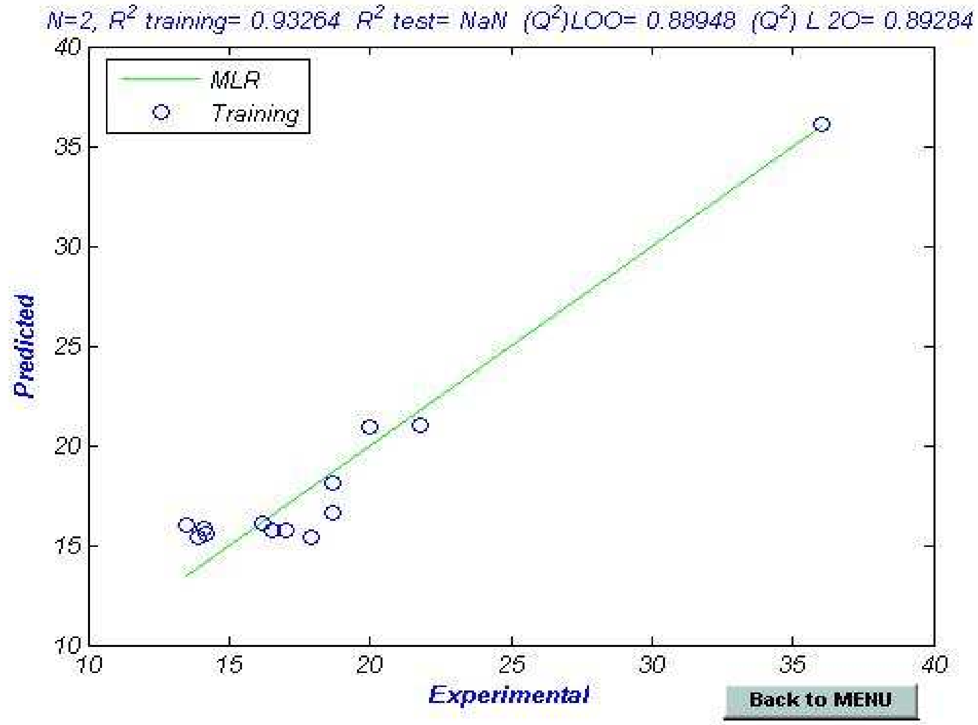


Figure1. The predicted versus the experimental pLD50 by MLR.

The real usefulness of QSAR models is not just their ability to reproduce known data, verified by their fitting power (R2), but is mainly their potential for predictive application. For this reason the model calculations were performed by maximising the explained variance in prediction, verified by the leave-one-out cross- the possibility of overestimating the model predictivity by using Q^2_{LOO} procedure, as is strongly recommended for QSAR modeling. The Q^2_{LOO} and Q^2_{LGO} for the MLR model are shown in Equation. This indicates that the obtained regression model has a good internal and external predictive power.

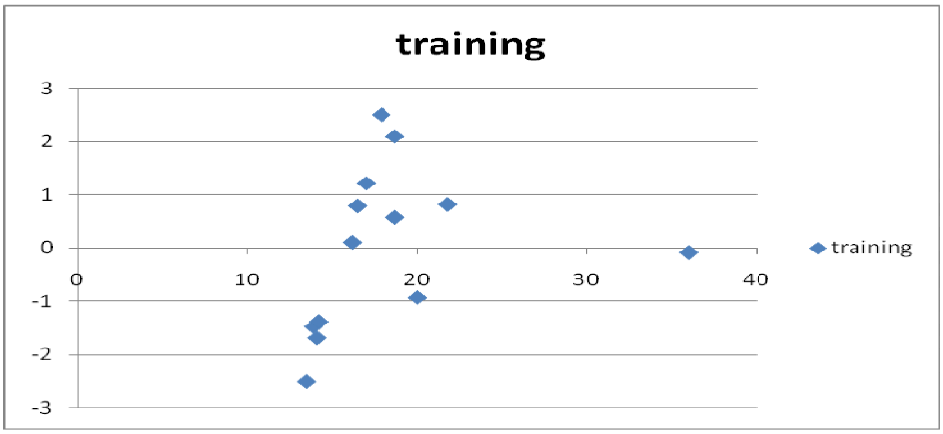


Figure 2. The residual versus the experimental pLD50 by GA-MLR. (See colour version of this figure online at www.informahealthcare.com/enz)

Also, in order to assess the robustness of the model, the Y-randomisation test was applied in this study. The dependent variable vector (pIC50) was randomly shuffled and a new QSAR model developed using the original independent variable matrix. The new QSAR models (after several repetitions) would be expected to have low R2 and Q^2_{LOO} values (Table 4). If the opposite happens then an acceptable QSAR model cannot be obtained for the specific modeling method and data.

Table 4. The R^2_{train} and Q^2_{LOO} values after several Y-randomisation tests

N	Q^2	R^2
1	0.000114	0.100605
2	0.004401	0.061712
3	0.170186	0.234025
4	0.244233	0.053246
5	0.236332	0.080586
6	0.001546	0.031958
7	8.97E-06	0.064639
8	0.81525	0.81598
9	0.055088	0.03324
10	0.051992	0.024975

Applicability domain

The Williams plot (Figure 3), the plot of the standardized residuals versus the leverage, was exploited to visualise the applicability domain . The leverage indicates a compound’s distance from the centroid of X. The leverage of a compound in the original variable space is defined as [24]:

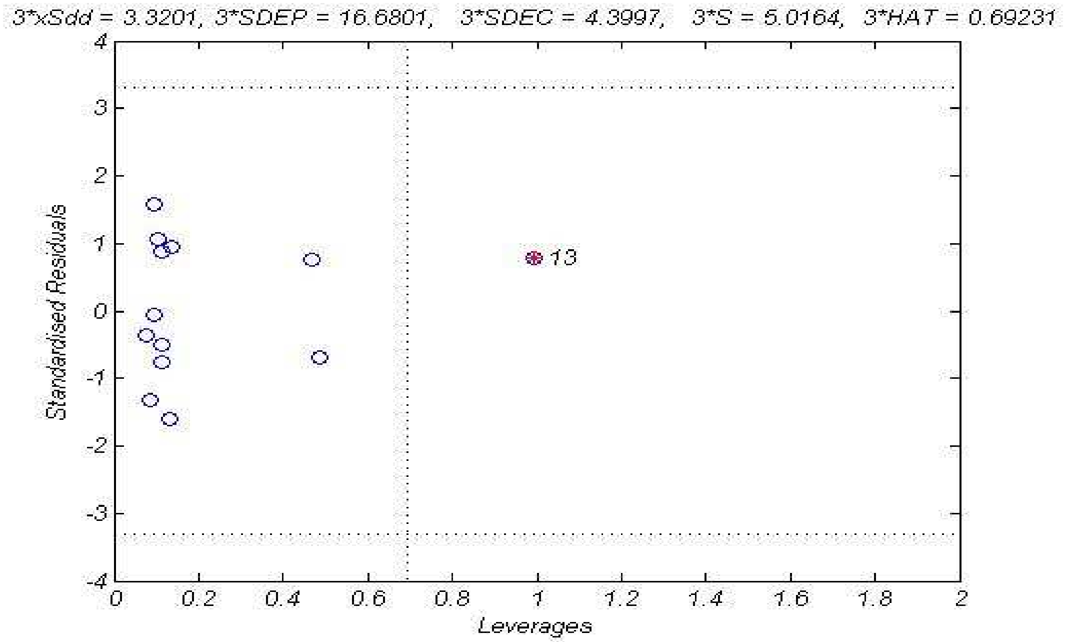


Figure 3. The William plot of the GA-MLR model.
(See colour version of this figure online at www.informahealthcare.com/enz)

$$hi=X_i^T(X_i^TX_i)^{-1} X_i \qquad (4)$$

Where *xi* is the descriptor vector of the considered compound and X is the descriptor matrix derived from the training set descriptor values. The warning leverage (h*) is defined as [25]:

$$H^* = 3(p+1)/n \qquad (5)$$

Where n is the number of training compounds, p is the number of predictor variables. A compound with hi > h* seriously influences the regression performance, but it doesn't appear to be an outlier because its standardized residual may be small, even though it has been excluded from the applicability domain. Moreover, a value of three for the standardised residual is commonly used as a cut-off value for accepting predictions, because points that lie ± 3 standardised residuals from the mean will cover 99% of normally distributed data [26]. Thus the leverage and the standardised residual were combined for the characterization of the applicability domain. From Figure 4, it is obvious that there are no outlier compounds with standard residuals >3δ for both the training and test sets. Also all the chemicals have a leverage lower than the warning h* value of 0.462

CONCLUSION

In this article, a QSAR study of 13 anti-cancer drugs was performed based on the theoretical molecular descriptors calculated by the DRAGON and GAUSSIAN software and selected. The built model was assessed comprehensively (internal and external validation) and all the validations indicated that the QSAR model built was robust and satisfactory, and that the selected descriptors could account for the structural features responsible for the anti-cancer drugs activity of the compounds. The QSAR model developed in this study can provide a useful tool to predict the activity of new compounds and also to design new compounds with high activity.

REFERENCES

[1] Sammes PG, Taylor JB. Comprehensive Medicinal Chemistry. Oxford: Pergamon Press, 1990:766.
[2] Riahi S, Pourbasheer E, Dinarvand R, Ganjali MR, Norouzi P. *Chem Biol Drug Des* **2008**; 74:165–172.
[3] Riahi S, Pourbasheer E, Ganjali MR, Norouzi P. *J Hazard Mater* **2009**; 166:853–859.
[4] Riahi S, Pourbasheer E, Ganjali MR, Norouzi P. *Chem Biol Drug Des* **2009**; 73:558–571.
[5] Depczynski U, Frost VJ, Molt K. *Anal Chim Acta* **2000**; 420:217.
[6] Alsberg BK, Marchand-Geneste N, King RD. *Chemometr Intel Lab* **2000**; 54:75–91.
[7] Jouanrimbaud D, Massart DL, Leardi R, Denoord OE. *Anal Chem* **1995**;67:4295–4301.
[8] Riahi S, Ganjali MR, E Pourbasheer, Divsar F, Norouzi P, Chaloosi M. *Curr Pharm Anal* **2008**; 4:231–237.
[9] Riahi S, Ganjali MR, Pourbasheer E, Norouzi P. *Chromatographia* **2008**; 67:917–922.
[10] Riahi S, Pourbasheer E, Ganjali MR, Norouzi P, Zeraatkar Moghaddam A. *J Chin Chem Soc* **2008**; 55:1086–1093.
[11] Riahi S, Ganjali MR, Moghaddam AB, Pourbasheer E, Norouzi P. *Curr Anal Chem* **2009**; 5:42–47.
[12] Riahi S, Pourbasheer E, Dinarvand R, Ganjali MR, Norouzi P. *Chem Biol Drug Des* **2008**; 72:205–216.
[13] Riahi S, Pourbasheer E, Dinarvand R, Ganjali MR, Norouzi P. *Chem Biol Drug Des* **2008**; 72:575–584.
[14] Monneret C. Eur. J. Med. Chem. **36**, 483-493(**2001**)
[15] http://www.chemidplus.com
[16] Holland H. Adaption in Natural and Artificial Systems. Ann Arbor, MI: The University of Michigan, **1975**; 342–375.
[17] Cartwright HM. Applications of Artificial Intelligence in Chemistry. Oxford: Oxford University, **1993**;760–765.
[18] Hunger J, Huttner G. Optimization and analysis of force field parameters by combination of genetic algorithms and neural networks. *J Comput Chem* **1999**; 20:455–471.
[19] Ahmad S, Gromiha MM. J *Comput Chem* **2003**; 24:1313–1320[20]. Waller CL, Bradley MP. *J Chem Inf Comput Sc*i **1999**; 39:345–355
[21] Aires-de-Sousa J, Hemmer MC, Casteiger J. Prediction of H-1 NMR chemical shifts using neural networks. *Anal Chem* **2002**; 74:80–90.
[22] The Mathworks. Genetic Algorithm and Direct Search Toolbox Users Guide. Massachusetts: MathWorks, **2002**; 50–65

[23] G. Espinosa, D. Yaffe, Y. Cohen, A. Arenas, F. Giralt, *J. Chem. Inf. Comput. Sc*i. **2000**, 40,859–879.

[24] Netzeva TI, Worth AP, Aldenberg T, Benigni R, Cronin MTD, Gramatica P, Jaworska JS, Kahn S, Klopman G, CA Marchant, Myatt G, Nikolova- Jeliazkova N, Patlewicz GY, Perkins R, Roberts DW, Schultz TW, Stanton DT, van de Sandt JJM, Tong W, Veith G, Yang C. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. ATLA-Altern Lab Anim **2005;** 33:155–173.

[25] Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM, Gramatica P. *Environ Health Perspect* **2003**; 111:1361–1375.

[26] Jaworska JS, Nikolova JN, Aldenberg T. *ATLA Altern Lab Anim* **2005**;33:445–459.