The application of several variable selection methods in the Predictive Toxicology Challenge 2000-2001.

T.H. Reijmers, M. Engels

Janssen Research Foundation
Molecular Design & Chemoinformatics – Beerse – Belgium

Introduction

The identification of potentially toxic compounds in the early phases of the pharmaceutical drug discovery process has become an appealing strategy (Durham and Pearl, 2001). However, toxicological studies are time and cost intensive and for that reason it has been difficult to address toxicological issues in the early discovery phase. In-silico toxicology approaches investigating relationships between chemical structures and their toxicological effects present a fast and inexpensive alternative to in-vivo experiments and have for that reason drawn quite some attention. However it has been questioned whether in-silico approaches have reached the degree of maturity and robustness that enables them to compete with the classical experiment.
The 2nd Predictive Toxicology Challenge (PTC) forms an interesting forum to test the predictive power of novel machine learning and statistical approaches, to evaluate the transparency of the resulting in-silico models, and to quantify their robustness (Helma et al., 2001). In this extended abstract we summarize our efforts to develop quantitative structure-activity relationships (QSAR) models for the 4 different sex/species data sets. In the Material and Methods section first the data set is introduced that is used during this study. Next the different methods are described that are evaluated for selecting an optimal set of descriptors. Finally we will briefly describe the prediction of the carcinogenic potential of a compound for the different sex/species models.

Material and Methods

Data

The data set that is used in the PTC consists of 417 chemicals with corresponding classifications. During the data engineering stage of the PTC over 1000 descriptors are submitted. In this study the 115 descriptors submitted by the authors are used only. Table 1 lists the submitted descriptors. Besides descriptors which characterize structural properties (CRIPPEN1-CRIPPEN72), several descriptors are used that describe certain physicochemical properties of the 417 chemicals. Furthermore density functional theory is used to calculate some electronic descriptors. The last rows of table 1 show the descriptors that are determined using the autocorrelation method with the atomic molecular orbital LUMO or HOMO coefficients as atomic property. The autocorrelation method is typically used to quantify the local distribution of a certain atomic property within the molecule. Finally, the whole set of descriptors is enriched with simple 1D descriptors providing simple counts of certain molecular properties.
In this study the NTP classifications P (positive), CE (clear evidence) and SE (some evidence) code for carcinogens, while N (negative), NE (no evidence), E (equivocal) and EE (equivocal evidence) code for non-carcinogens. Chemicals with a missing

classification and with the classification IS (inadequate study) are excluded from further calculations.

Methods

In order to extract the most relevant properties for model building a two stage procedure is applied. In the first stage, frequency and correlation analyses are applied to all descriptors. Descriptors are dropped either when they are not sufficiently represented in the data set (< 5 occurrences) or the pairwise correlation is higher than 0.9. Of the 72 atom type descriptors according to Wildman & Crippen, 22 are dropped due to the frequency analyses. During the correlation analyses, 4 Wildman & Crippen descriptors and 7 other descriptors are removed (CMR, SLOGP, SMR, HBA, LOGD (pH=7.4), VOLUME and EL_NEG). By means of these analyses the initial number of descriptors is reduced from 115 to 82.

Table 1. The initial set of 115 descriptors.

Descriptors		
1-72	CRIPPEN 1 – CRIPPEN 72	Atom type descriptors according to Wildman and Crippen, 1999.
73	CLOGP	Calculated logP using Daylight program.
74	CMR	Calculated molar refractivity using Daylight program.
75	SLOGP	Calculated logP according to Wildman and Crippen, 1999.
76	SMR	Calculated molar refractivity according to Wildman and Crippen, 1999.
77	ROTBOND	Number of single, non-cyclic bonds.
78	FLEX	Degree of flexibility (0 rigid, 1 extremely flexible).
79	TPSA	Topological polar surface area according to Ertl et al., 2000.
80	MW	Molecular weight.
81	HBD	Number of H-bond donors.
82	HBA	Number of H-bond acceptors.
83–85	LOGD (2, 7_4, 10)	Calculated logD at pH=2.0, 7.4 and 10.0.
86	VOLUME	Volume based on CORINA generated conformations.
87	SURF_AREA	Surface area based on CORINA generated conformations.
88	DISTANCE	VOLUME / SURF_AREA.
89	HOMO	HOMO energy based on CORINA generated conformations, using density functional theory.
90	LUMO	LUMO energy based on CORINA generated conformations, using density functional theory.
91	DIPOLE	Dipole based on CORINA generated conformations, using density functional theory.
92	HARDNESS	Calculated hardness using HOMO and LUMO energies.
93	SOFTNESS	Calculated softness using HOMO and LUMO energies.
94	EL_NEG	Calculated electronegativity using HOMO and LUMO energies.
95	EL_PHIL	Calculated electrophilicity using HOMO and LUMO energies.
96-100	ATOM_LUMO1	Autocorrelation vector using atomic molecular orbital coefficients of LUMO.
101-105	ATOM_HOMO1	Autocorrelation vector using atomic molecular orbital coefficients of HOMO.
106-110	ATOM_LUMO2	ATOM_LUMO1 × LUMO.
111-115	ATOM_HOMO2	ATOM_HOMO1 × HOMO.

Up till now descriptors are dropped solely on the basis of information in the descriptor space (the input variable space). In the second stage of the variable selection procedure, several variable selection methods are applied that additionally use the carcinogenicity classifications (output/target values). Ultimately this will result for each of the 4 different sex/species data sets in 4 optimized sets of descriptors that will be used in the final modeling.

The first applied variable selection method (the R^2-method) starts by calculating the squared correlation coefficient (R^2) between each descriptor and the classification results. Descriptors with a R^2 less than a cutoff criterion (0.005) are dropped. The remaining descriptors are used in a forward stepwise regression. First the descriptor with the highest R^2 is selected to be used as input for the regression. Next descriptors are added to the input until no significant improvement of the regression model occurs. The descriptors that have been kept to perform the regression analysis are considered to be the most significant descriptors and will be used to construct the final

model. In the χ^2-method the relationship between a descriptor and the classification is examined by first decomposing each descriptor into several binary dummy variables. Next two-way frequency tables are created (binary descriptor against binary classification). The χ^2-test is used to examine if the observed frequencies differ significantly from the expected frequencies. If the observed frequencies differ from the expected frequencies the descriptor is considered relevant and will be used to construct the final model. The last two variable selection methods are very similar. In both applications a global optimization algorithm (a genetic algorithm (GA)) is used to find an optimal subset of descriptors that gives the best classification results for a given modeling technique. In the GA-PLS variable selection method partial least squares regression (PLS) is used to model the relationship between carcinogenicity and structure information. The GA-LR method uses logistic regression as modeling technique. To prevent the modeling techniques from overfitting, in both cases cross-validation is applied.

Table 2. Results of four different variable selection methods on the male rat data set.

	Variable Selection Method			
	R^2	χ^2	GA-PLS	GA-LR
Variables				
CRIPPEN 1	×	×		
CRIPPEN 2	×			×
CRIPPEN 5	×			
CRIPPEN 11	×		×	×
CRIPPEN 12	×		×	×
CRIPPEN 21				×
CRIPPEN 22	×		×	×
CRIPPEN 29	×			
CRIPPEN 30	×			
CRIPPEN 34	×			
CRIPPEN 35	×			
CRIPPEN 36	×		×	
CRIPPEN 37	×		×	×
CRIPPEN 39	×		×	
CRIPPEN 41				×
CRIPPEN 49			×	×
CRIPPEN 54			×	×
CRIPPEN 57	×			
CRIPPEN 58			×	
CRIPPEN 63			×	
CRIPPEN 64	×	×	×	×
CRIPPEN 67				×
CRIPPEN 68	×			

Finally for each sex/species target, the optimized set of descriptors is used to train a backpropagation neural network (NN). This results in four different models predicting carcinogenicity for the male rat (MR), the female rat (FR), the male mouse (MM) and

the female mouse (FM). All calculations are performed using Matlab and SAS EnterpriseMiner software.

Results & Discussion

Because for most modeling techniques the number of descriptors they can deal with is limited, significant descriptors (variables) have to be selected prior to modeling. The male rat classifications are used to examine the efficiency of different variable selection methods. In table 2 the results of the four variable selection methods on the Wildman & Crippen descriptors are shown. For clarity reasons the selection results of the other descriptors are left out. While the R^2-method and the GA-based methods each select more than 10 variables, the χ^2-method selects 2 descriptors only: the atom type descriptors corresponding to the total number of 1° and 2° aliphatic carbons and bromines respectively. Indeed almost all chemicals containing bromine are carcinogens causing an unbalanced frequency table and a negative χ^2-test. Not only the χ^2-method selects this descriptor but all selection methods recognize that this descriptor is important for modeling. When the table is further examined we see that the selected variables of the R^2-method and the GA-based methods are very similar. By means of the variable selection methods the number of significant descriptors could be reduced to about 15% of the original descriptor set.
To see how these different variable selection methods influence the predictions of the classifications models both logistic regression (LR) and neural network (NN) models are constructed using the optimized descriptor sets. In table 3 the results are visualized.

Table 3. Male rat concordances of the different logistic regression and neural network models created on the basis of descriptor sets generated by the different variable selection methods.

		Concordance (%)	
variable selection method	modeling technique	training	validation
	LR		
R^2		68	57
χ^2		62	52
GA-PLS		69	54
GA-LR		69	58
	NN		
R^2		76	60
χ^2		64	58
GA-PLS		74	62
GA-LR		73	58

Although no variable selection methods are used that are capable of discovering significant non-linear interactions between the descriptors and the classifications, for all optimized descriptor sets the NN gives better results than the LR. Furthermore for each model the concordance of the validation set (30% of the available objects) is considerably lower than the concordance of the training set (70 % of the available objects). On the basis of these concordance results no clear preference can be given for one of the variable selection methods. There is no selection method that performs significant better than the other methods. At the very most, the worse performance of the χ^2-method in combination with the NN is noticed.

Table 4 shows the selected descriptors by means of GA-PLS for the male and the female rat/mouse classifications (MR, FR, MM and FM respectively).

Table 4.	Selected descriptors by GA-PLS for the different species/sex models.
model	*selected descriptors*
MR	CRIPPEN: 11, 12, 22, 36, 37, 39, 49, 54, 58, 63, 64 MW; ATOM_LUMO1; ATOM_HOMO1; ATOM_HOMO2
FR	CRIPPEN: 3, 11, 12, 22, 30, 34, 37, 39, 51, 54, 63, 64 MW; LOGD (pH=10); DIPOLE; ATOM_HOMO1; ATOM_HOMO2
MM	CRIPPEN: 1, 8, 11, 25, 30, 34, 36, 49, 54, 55, 59, 67 TPSA; MW; HARDNESS; ATOM_LUMO1; ATOM_HOMO1; ATOM_HOMO2
FM	CRIPPEN: 3, 23, 30, 34, 37, 40, 49, 68 DISTANCE; DIPOLE; ATOM_HOMO2

Because the classification results for MR and FR are highly correlated, the GA-PLS method has selected similar descriptors (for the most part structural and electronic descriptors) to create the two different rat models. The selected descriptors for the mouse models are far less similar. This is probably caused by the relative large number of compounds in the data set that have carcinogenicity classifications for the male or female mouse only but no classifications for both sexes.
Because in table 3 the best results are obtained with a NN, this modeling technique is also used to create the final models for the different sex/species carcinogenicity classifications. Again 70 % of the available data is used to create the model and 30 % is used for validation purposes. Concordances of 62% and 74% for male and female rat indicate that the NN has difficulties modeling carcinogenicity. Concordances of about 60 % for the male and female mouse confirm this. Although the number of descriptors is significantly reduced, no satisfactory results have been obtained accept for the female rat.

Conclusion

The total number of available descriptors that can be used to model structure activity relationships is growing constantly. Because most modeling techniques are designed to deal with a limited number of descriptors and it is not always clear which descriptors should be used for certain applications, variable selection methods, which search for the optimal subset of descriptors, are necessary. In the PTC also a large number of descriptors are available and as a consequence several different variable selection methods are applied and evaluated. Although different variables are selected by the different methods, on the basis of the performances of the corresponding models no clear preference can be given for one of the methods.

References

Durham, S.K. and Pearl, G.M., **2001**. Computational methods to predict drug safety liabilities. *Current Opinion in Drug Discovery & Development* 4, 110-115.

Ertl, P., Rohde, B. and Selzer, P., **2000**. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of Medicinal Chemistry* 43, 3714-3717.

Helma, C., King, R.D., Kramer, S. and Srinivasan, A., **2001**. The predictive toxicology challenge 2000-2001. *Bioinformatics* 17, 107-108.

Wildman, S.A. and Crippen, G.M., **1999**. Prediction of physicochemical parameters by atomic contributions. *Journal of Chemical Information and Computer Sciences* 39, 868-873.