QSAR of carcinogenicity and mutagenicity using codes of cycles in optimal SMILES-based descriptors

A. Chana, A.A. Toropov, A.P. Toropova and E. Benfenati

Istituto di Ricerche Farmacologiche Mario Negri, 20156, Via La Masa 19, Milano, Italy

Abstract

Carcinogenicity is an important endpoint for REACH, and typically for this endpoint many animals are used. Some in silico models exist, which in most of the cases are aimed to classify chemicals as carcinogenic or not. REACH requires an evaluation of the risk in case of the use of carcinogenic compounds, considering the exposure levels. For this, QSAR models, predicting a potency level, and not classifiers, may play a role. We developed QSAR models based on simplified molecular input line entry system (SMILES). SMILES has been used as elucidation of the molecular structure for quantitative structure – activity relationships aimed to predict carcinogenicity of large dataset that contains wide variety of organic compounds. Using the Monte Carlo method we constructed optimal descriptors, which are a mathematical function of composition of the SMILES elements together with special codes of cycles present in molecules. The codes of cycles indicate the presence of: cycles with sizes 5 and 6, cycles with hetero-atoms and condensed cycles. We will show that taking into account of the codes of cycles improves the predictive ability of the optimal descriptors for the external test set. In addition this approach has been used to model mutagenicity of heteroaromatic amines.

Briefly about SMILES notation

Optimal descriptors calculated with SMILES have been used for quantitative structure – property/activity relationships (QSPR/QSAR) [1-3]. In case of the optimal descriptors calculated with molecular graph (hydrogen filled) the statistical characteristics of the models improve if information on cycles is added. Similar approach based on the SMILES-based optimal descriptors has indicated that statistical characteristics of the QSAR for carcinogenicity are also preferable The technique of the blocking rare SMILES attributes has been used. The discrimination of the SMILES attributes into rare and not rare was carried out with a special threshold limS. limS is the minimal number of a SMILES attribute in the training set. If less then limS SMILES contain the attribute SAk*, than CW(SAk*)=0.0, i.e., the SAk* has no influence on the model.

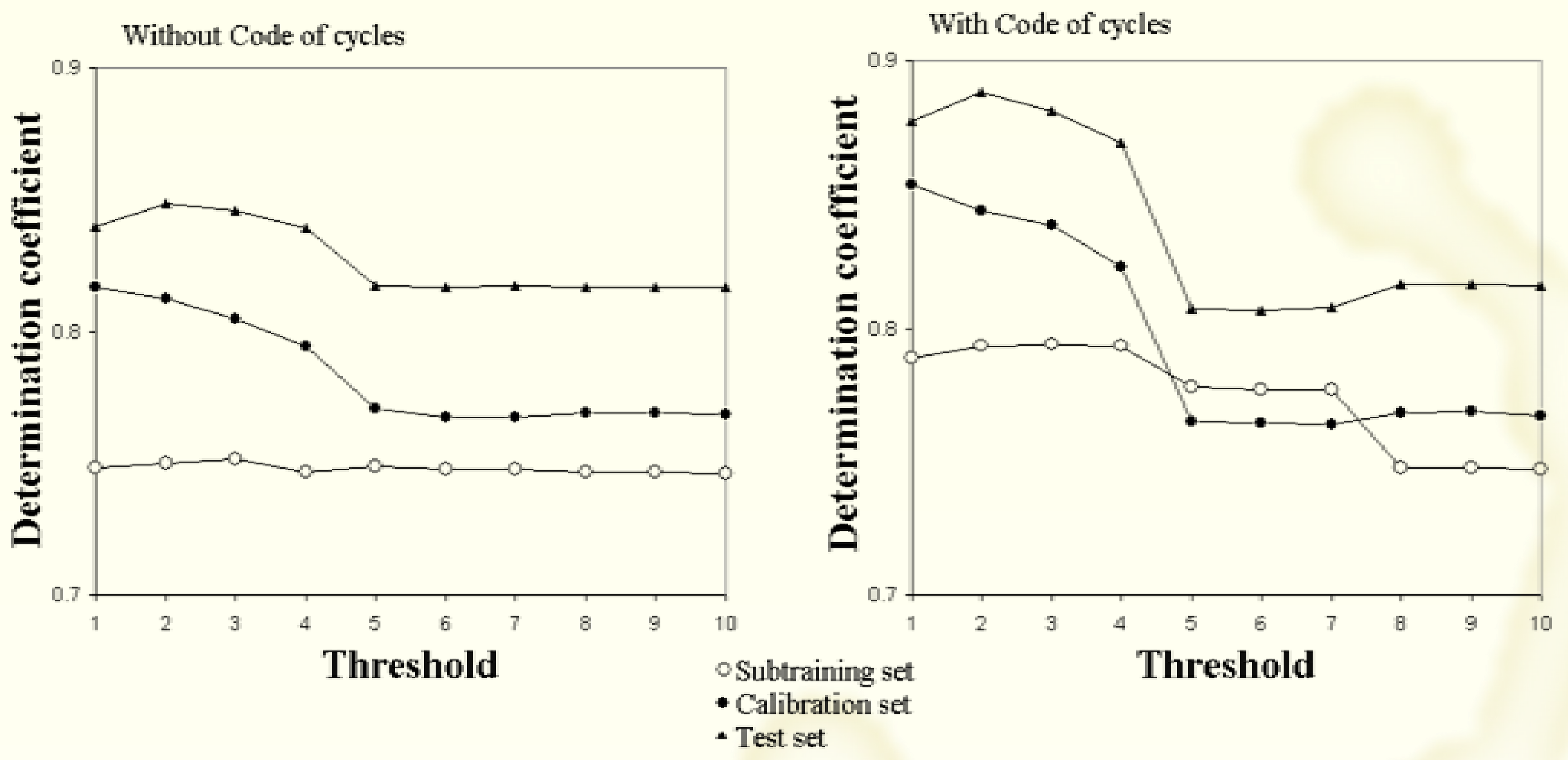


Figure 1

Statistical quality of the modelsof mutagenicity (logR) for different values of the threshold, LimS

Method

Two versions of the SMILES-based optimal descriptors have been studied:

1. without of the cycles code

$$\mathbf{DCW(limS) = \Sigma\, CW(SAk)} \qquad \mathbf{(1)}$$

2. with cycle codes

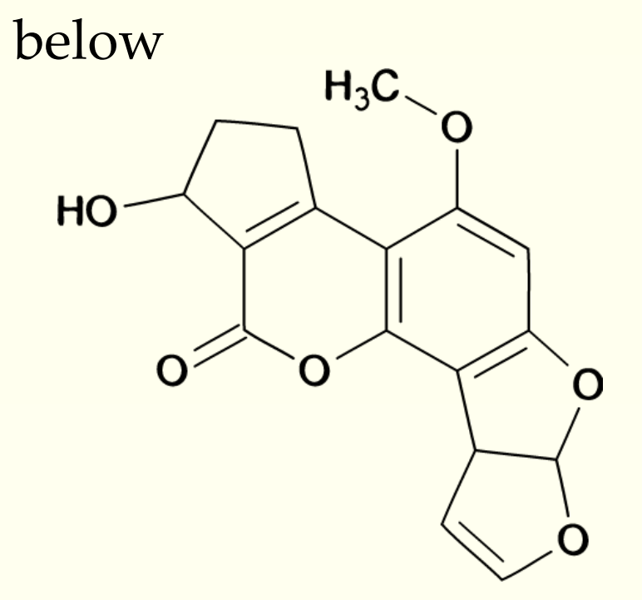
$$\mathbf{DCW(limS) = CW(CC) + \Sigma\, CW(SAk)} \qquad \mathbf{(2)}$$

where SAk are the SMILES attributes constructed with three consequent SMILES elements (i.e., one symbol, or two symbols which can not be examined separately , e.g., 'Cl', 'Br'; dC is difference of number of carbon atoms in sp2 state minus number of carbon atoms in sp3 state; CC is the cycles code for a given SMILES. CW(x) is the correlation weight for x (x is a SMILES attribute).

Cycles codes have been defined as the following

&(5-member cycles number)(6-member cycles number)(heteroatoms number)

For example, the compound below



can be represent by the SMILES: **O=C2Oc1c4C5C=COC5Oc4cc(OC)c1C=3CCC(O)C2=3**

The cycle code for the compound is **&321**

Rings have been calculated with the algorithm from Ref. 4, since from the SMILES code it is impossible to extract the full set of rings in macrocyclic condensed systems being such structures non explicitly expressed in the code. Therefore we decided to extract the adjacency matrix from the SMILES code and determine the total number of cycles, and their characteristics, present within every molecule. Cycles are classified in size, number of occurrences and heteroatomic content, classification that will be expressed ultimately in the cyclicity invariant code.

Best models for the *carcinogenicity* (-pTD50) and *mutagenicity* TA98 (logR)

Statistical characteristics of models for carcinogenicity (with code of cycle)

		Subtraining set				Calibration set				Test set			
Threshold	**Nact**	**N**	**r^2**	**s**	**F**	**n**	**r2**	**s**	**F**	**n**	**r2**	**s**	**F**
1	601	170	0.7840	0.659	610	170	0.7833	0.686	607	61	0.6417	0.784	106
2	397	170	0.7029	0.773	397	170	0.7042	0.774	400	61	0.7214	0.669	153
3	315	170	0.6688	0.816	339	170	0.6679	0.827	338	61	0.7198	0.665	152
4	276	170	0.6444	0.845	304	170	0.6498	0.846	312	61	0.7356	0.655	164
5	239	170	0.6124	0.883	265	170	0.6313	0.865	288	61	**0.7729**	**0.599**	201
6	210	170	0.5790	0.920	231	170	0.5894	0.908	241	61	0.7078	0.676	143
7	187	170	0.5623	0.938	216	170	0.5751	0.922	227	61	0.7326	0.654	162
8	161	170	0.5410	0.960	198	170	0.5708	0.927	223	61	0.7097	0.671	144
9	148	170	0.4993	1.003	168	170	0.5465	0.958	202	61	0.6765	0.709	123
10	139	170	0.5180	0.984	181	170	0.5633	0.941	217	61	0.6824	0.702	127

The statistical characteristics of the previous model (without the code of cycles) for *carcinogenicity* [1] are n=170, r2= 0.75, s=0.71 (subtraining set); n=170, r2=0.75, s=0.68 (calibration set); n=61, r2=0.72, s=0.70 (test set)

References

[1] A. A. Toropov, A.P. Toropova, E. Benfenati, Mol. Divers. *In press*
[2] A.A. Toropov, E. Benfenati, Cur. Drug Disc. Tech., 4 (2007) 77-116
[3] A. A. Toropov, E. Benfenati, Bioorg. Med. Chem. 16 (2008) 4801–4809
[4] Th. Hanser, Ph. Jauffret, G. Kaufmann J. Chem. Inf. Comput. Sci. 36(1996) 1146-1152

Statistical characteristics of models for mutagenicity

		Subtraining set				Calibration set				Test set			
Without code of cycle													
Threshold	**Nact**	**N**	**r^2**	**s**	**F**	**n**	**r2**	**s**	**F**	**n**	**r2**	**s**	**F**
1	18	42	0.7482	1.102	119	25	0.8168	0.738	103	28	0.8396	0.722	136
2	17	42	0.7499	1.098	120	25	0.8122	0.751	100	28	**0.8485**	**0.703**	146
3	16	42	0.7514	1.094	121	25	0.8044	0.776	95	28	0.8458	0.713	143
4	14	42	0.7464	1.105	118	25	0.7943	0.780	89	28	0.8390	0.729	135
5	12	42	0.7488	1.100	119	25	0.7708	0.822	77	28	0.8172	0.787	116
6	8	42	0.7474	1.103	118	25	0.7675	0.829	76	28	0.8169	0.786	116
7	8	42	0.7477	1.103	119	25	0.7676	0.831	76	28	0.8172	0.785	116
8	8	42	0.7467	1.105	118	25	0.7689	0.826	77	28	0.8168	0.786	116
9	8	42	0.7468	1.104	118	25	0.7689	0.827	77	28	0.8168	0.787	116
10	7	42	0.7460	1.106	117	25	0.7685	0.832	76	28	0.8167	0.790	116
With code of cycle													
Threshold	**Nact**	**N**	**r^2**	**s**	**F**	**n**	**r2**	**s**	**F**	**n**	**r2**	**s**	**F**
1	24	42	0.7886	1.009	149	25	0.8535	0.660	134	28	0.8771	0.629	186
2	22	42	0.7935	0.998	154	25	0.8442	0.680	125	28	**0.8880**	**0.603**	206
3	21	42	0.7936	0.997	154	25	0.8383	0.704	119	28	0.8809	0.623	192
4	18	42	0.7935	0.998	154	25	0.8227	0.739	107	28	0.8692	0.658	173
5	16	42	0.7778	1.035	140	25	0.7646	0.843	75	28	0.8074	0.839	109
6	12	42	0.7764	1.038	139	25	0.7639	0.841	74	28	0.8069	0.836	109
7	12	42	0.7766	1.037	139	25	0.7637	0.843	74	28	0.8077	0.835	109
8	10	42	0.7474	1.103	118	25	0.7679	0.828	76	28	0.8165	0.787	116
9	10	42	0.7472	1.104	118	25	0.7684	0.828	76	28	0.8163	0.788	116
10	9	42	0.7468	1.105	118	25	0.7671	0.835	76	28	0.8160	0.792	115

Conclusions

The code of cycle has improved the predictive potential of both QSAR-models

for the carcinogenicity and for the mutagenicity

Acknowledgements

The authors thank the *Marie Curie Fellowships,* through the contract MIF1-CT-2006-039036 - **CHEMPREDICT**, and the EC funded project **CAESAR** (contract SSPI-022674) for financial support