QSAR MODELING OF CARCINOGENICITY BASED ON LOCAL ATTRIBUTES OF SMILES AND SPECIAL CODES OF CYCLES (GLOBAL SMILES ATTRIBUTES)

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

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

Introduction

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 SMILES. Simplified molecular input line entry system (SMILES) has been used as elucidation of the molecular structure for quantitative structure – activity relationships which are 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 are reflected a presence of: cycles with sizes 5 and 6, cycles with hetero-atoms and condensed cycles.

Materials & Methods

Optimal descriptors calculated with simplified molecular input line entry system (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) statistical characteristics of the models becomes better if their calculating includes information on cycles. Similar approach based on the SMILES-based optimal descriptors has indicated that statistical characteristics of the QSAR for carcinogenicity are also preferable (Table 1). Technique of blocking of 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 than limS SMILES contain the attribute SAk*, than CW(SAk*)=0.0, i.e., the SAk* has no influence to the model.

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

1. without cycle codes

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

2. with cycle codes

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

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 cycle code for a given SMILES. CW(x) is the correlation weight for x (x is a SMILES attribute).

Results & Discussions

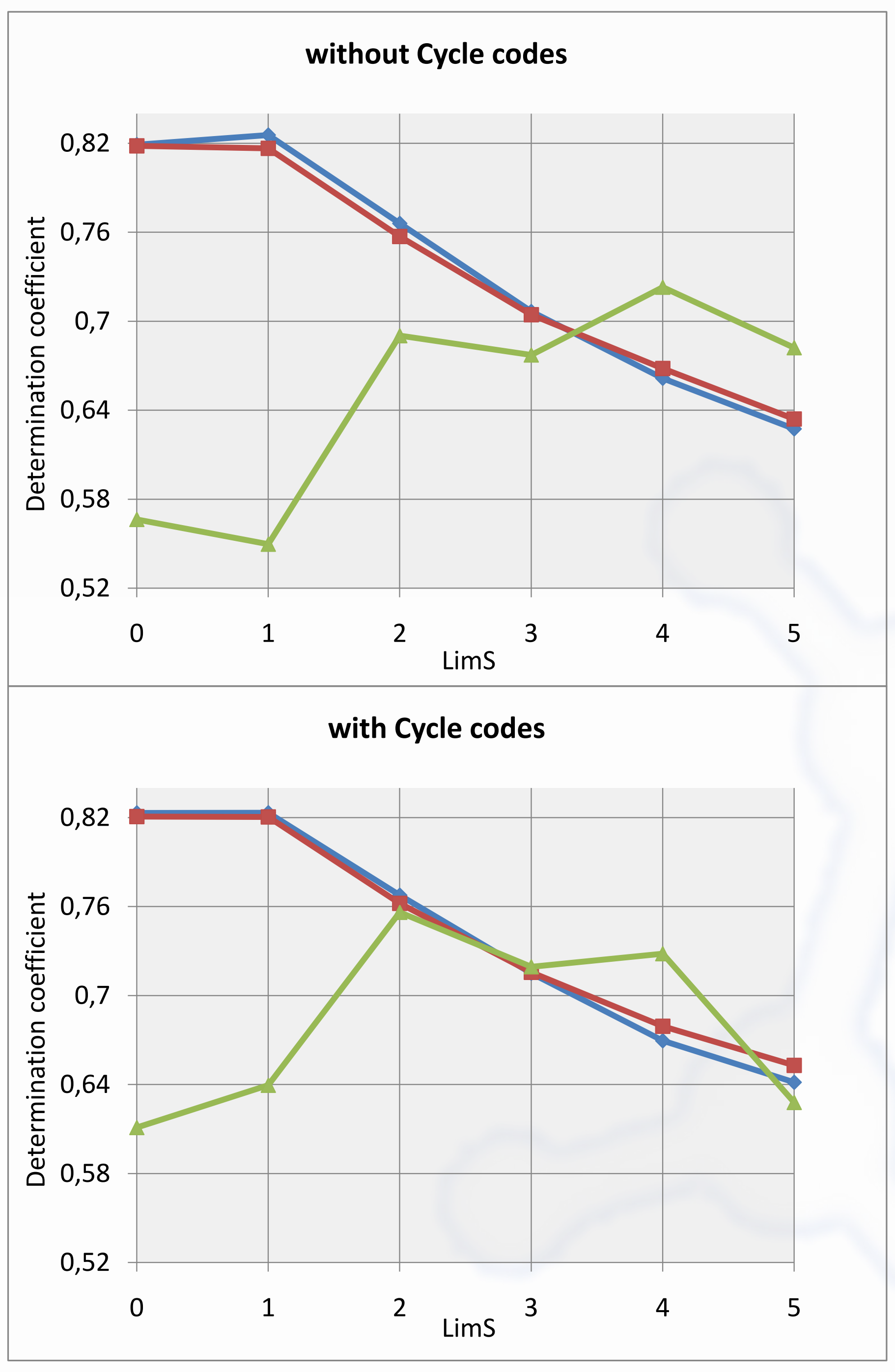


Figure 1
Statistical quality of the models for the training (red line), calibration (blue line), and test (green line): the cases of the Monte Carlo optimization with Eq. 1 and 2 obtained on range of the limS of 0-5.

Table 1
Statistical characteristics of the Models, for different limN and two versions of the descriptors.

Models obtained WITHOUT Cycle codes

			Training set, n=170			Calibration set, n=170			Test set, n=61		
imNI	Nact	Probe	r2	s	F	r2	s	F	r2	s	F
0	593	1	0.8168	0.607	749	0.8133	0.622	732	0.5259	0.971	65
0	593	2	0.8211	0.600	771	0.8149	0.617	740	0.5386	0.968	69
0	593	3	0.8162	0.608	746	0.8156	0.625	743	0.4969	1.026	58
0			0.8180	0.605	755	0.8146	0.622	738	0.5205	0.988	64
1	469	1	0.8197	0.602	764	0.8176	0.630	753	0.5577	0.988	74
1	469	2	0.8173	0.606	751	0.8132	0.635	732	0.5960	0.901	87
1	469	3	0.8193	0.603	762	0.8147	0.645	738	0.6088	0.891	92
1			0.8188	0.603	759	0.8152	0.636	741	0.5875	0.927	84
2	311	1	0.7643	0.688	545	0.7557	0.699	520	0.6847	0.760	128
2	311	2	0.7648	0.687	546	0.7651	0.686	547	0.6932	0.784	133
2	311	3	0.7678	0.683	556	0.7614	0.691	536	0.6938	0.773	134
2			0.7656	0.686	549	0.7608	0.692	534	**0.6906**	0.772	132
3	240	1	0.7120	0.761	415	0.7111	0.764	414	0.6579	0.790	113
3	240	2	0.7105	0.763	412	0.7093	0.766	410	0.6917	0.736	132
3	240	3	0.7093	0.764	410	0.7090	0.770	409	0.6674	0.758	118
3			0.7106	0.763	413	0.7098	0.767	411	0.6723	0.761	121
4	205	1	0.6628	0.823	330	0.6678	0.815	338	0.7377	0.641	166
4	205	2	0.6669	0.818	336	0.6691	0.813	340	0.7083	0.673	143
4	205	3	0.6677	0.817	338	0.6681	0.817	338	0.7227	0.657	154
4			0.6658	0.819	335	0.6683	0.815	339	0.7229	0.657	154
5	176	1	0.6358	0.855	293	0.6407	0.851	300	0.6775	0.725	124
5	176	2	0.6337	0.858	291	0.6349	0.857	292	0.6812	0.721	126
5	176	3	0.6253	0.868	280	0.6296	0.863	286	0.6795	0.713	125
5			0.6316	0.860	288	0.6351	0.857	292	0.6794	0.720	125

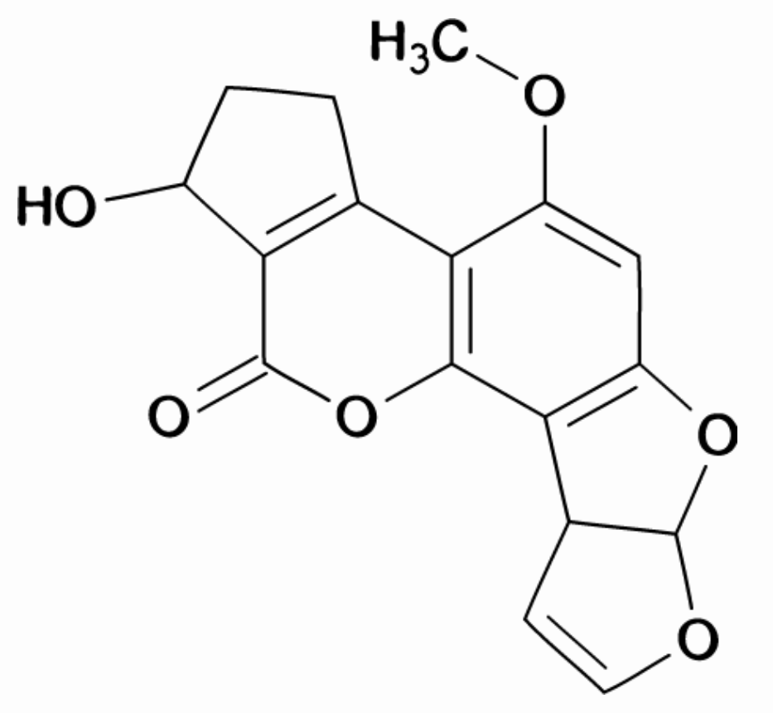
Models obtained WITH Cycle codes

			Training set, n=170			Calibration set, n=170			Test set, n=61		
imNI	Nact	Probe	r2	s	F	r2	s	F	r2	s	F
0	593	1	0.8198	0.602	764	0.8195	0.603	763	0.6050	0.848	90
0	593	2	0.8251	0.593	793	0.8209	0.606	770	0.6069	0.860	91
0	593	3	0.8245	0.594	789	0.8216	0.610	774	0.6212	0.809	97
0			0.8231	0.596	782	0.8207	0.606	769	0.6110	0.839	93
1	469	1	0.8240	0.595	787	0.8196	0.637	763	0.6504	0.818	110
1	469	2	0.8230	0.596	781	0.8208	0.636	770	0.6531	0.825	111
1	469	3	0.8226	0.597	779	0.8210	0.625	771	0.6152	0.884	94
1			0.8232	0.596	782	0.8205	0.633	768	0.6395	0.842	105
2	311	1	0.7699	0.680	562	0.7591	0.696	529	0.7615	0.656	188
2	311	2	0.7682	0.682	557	0.7646	0.686	546	0.7492	0.671	176
2	311	3	0.7647	0.688	546	0.7630	0.692	541	0.7577	0.668	185
2			0.7676	0.683	555	0.7622	0.692	539	**0.7561**	0.665	183
3	240	1	0.7123	0.760	416	0.7140	0.758	419	0.7145	0.699	148
3	240	2	0.7199	0.750	432	0.7199	0.754	432	0.7259	0.683	156
3	240	3	0.7129	0.760	417	0.7132	0.759	418	0.7179	0.677	150
3			0.7150	0.757	422	0.7157	0.757	423	0.7194	0.686	151
4	205	1	0.6734	0.810	346	0.6784	0.806	354	0.7370	0.635	165
4	205	2	0.6731	0.810	346	0.6762	0.807	351	0.7174	0.658	150
4	205	3	0.6621	0.824	329	0.6834	0.802	363	0.7301	0.644	160
4			0.6695	0.815	341	0.6793	0.805	356	0.7282	0.646	158
5	176	1	0.6466	0.843	307	0.6516	0.838	314	0.6019	0.806	89
5	176	2	0.6402	0.850	299	0.6601	0.831	326	0.6328	0.765	102
5	176	3	0.6379	0.853	296	0.6471	0.844	308	0.6491	0.744	109
5			0.6415	0.849	301	0.6529	0.838	316	0.6279	0.772	100

Cycle codes have been defined as the following
&(5-member cycles number)(6-member cycles number)(heteroatoms number)

The compound in *Figure 2* can be represented 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. 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. Results from *Table 1* and *Figure 1* show good prediction on the test set.

Figure 2
Example of a compound of view



Conclusions

One can see from *Table 1* and *Figure 1* that:

i) better results for both schemes take place if the limS=2;

ii) the model that involves cycles codes gives better prediction for the carcinogenicity of external test set.

References

[1] A. A. Toropov, E. Benfenati, Eur. J. Med. Chem. 42 (2007) 606-613

[2] 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

Acknowledgements

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