

QSAR modelling of carcinogenicity and mutagenicity potency by optimal SMILES-based descriptors

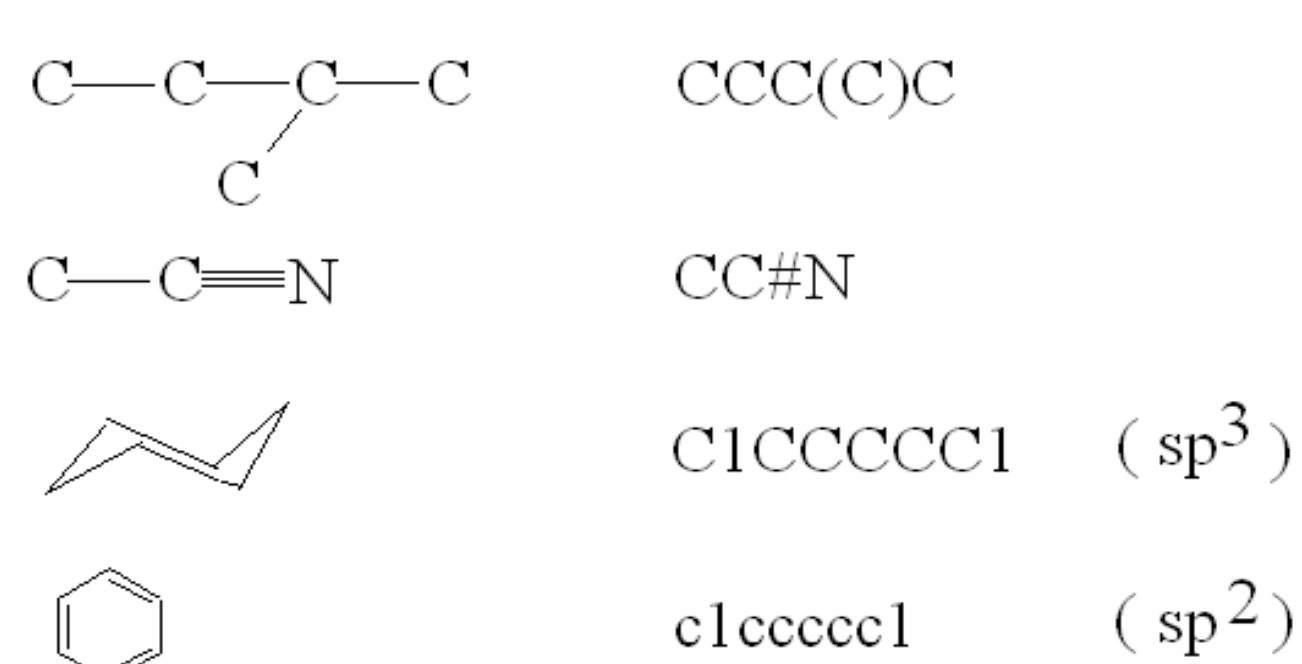
A. A. Toropov, A. P. Toropova and E. Benfenati

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

Abstract

Simple models for carcinogenicity and mutagenicity prediction are presented. The chemical information is based on descriptors extracted from the SMILES. These descriptors are used to build up regression equations, for both mutagenicity and carcinogenicity. A possibility of definition of applicability domain for the examined models is discussed.

Briefly about SMILES notation



<http://www.daylight.com/smiles/index.html>

Method

Data. Experimental values of the carcinogenicity is expressed as potency to induce cancer (C, in mmol/kg of body weight). As endpoint for modelling the $\log_{10}(1/C)$ has been used. 401 substances with C values for rat have been used. The split into training (n=360) and test (n=61) sets is random, but range of variation of the C values for these sets is almost identical.

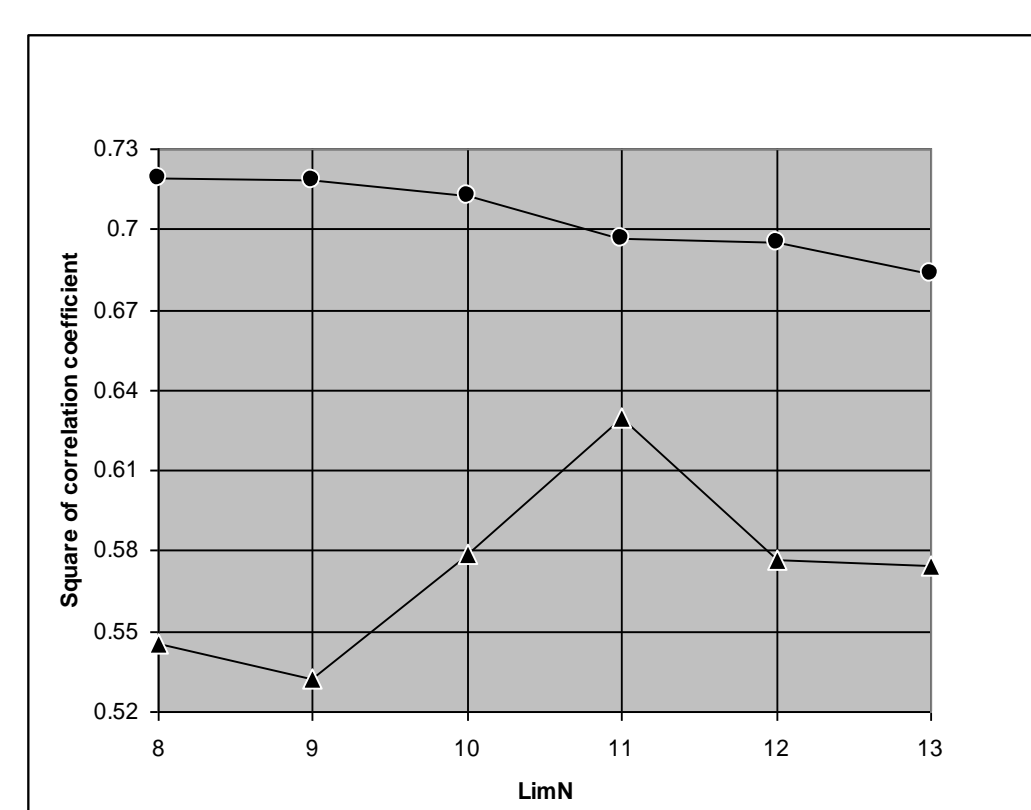
Experimental data on mutagenicity is expressed in $\log(\text{revertant/nmol})$, *S. typhimurium* TA98+S9. The SMILES-based optimal descriptors of the correlation weights (DCW) have been calculated as

$$DCW = \Pi CW(sk) \Pi CW(ssk) \Pi CW(sssk) \quad (1)$$

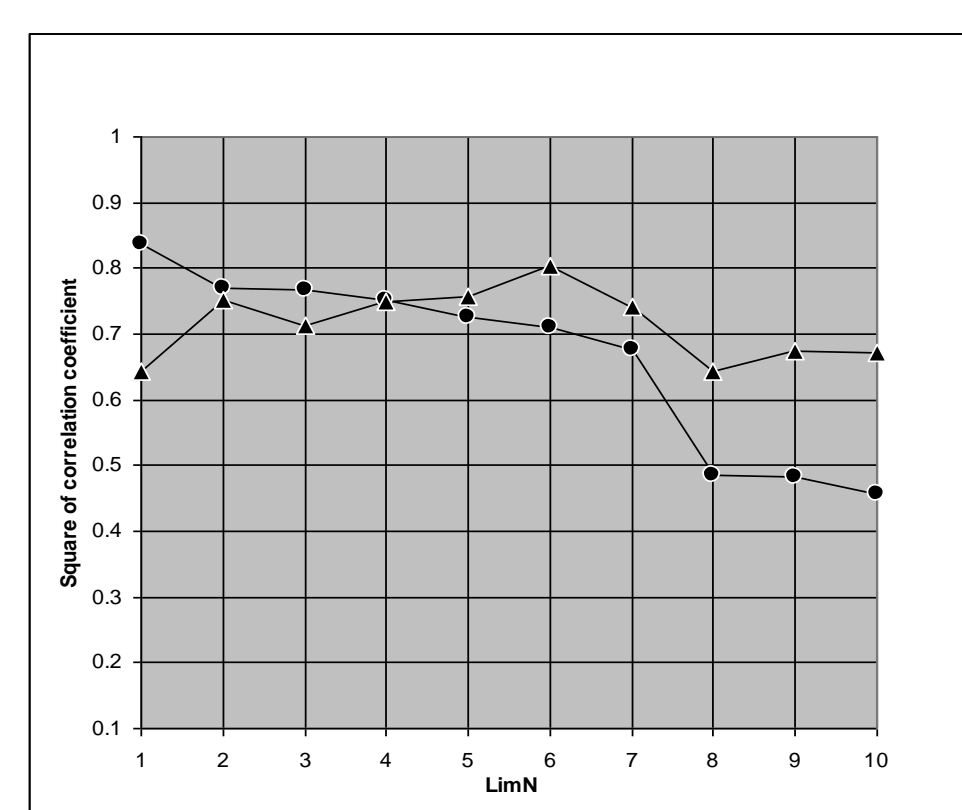
where sk , ssk , and $sssk$ are SMILES attributes (SAk) of one, two, and three elements, respectively. The element of the SMILES can be a symbol of the SMILES notation (for instance, 'c', 'C', 'n', 'N', '=', etc.), or two symbols encode an image (for instance, 'Cl', 'Br', 'N+', etc.); $CW(x)$ is correlation weight for the SMILES attribute x. The CWs are calculating by the Monte Carlo method optimization procedure that provides values which, used in equation (1), give maximal separation of the SMILES notation in the sk , ssk , and $sssk$ attributes and extended connectivity of zero- (vertex), first- (edge), and second order (path of length 2) defined in molecular graph.

Rare SMILES attributes leads to overfitting

Carcinogenicity



Mutagenicity

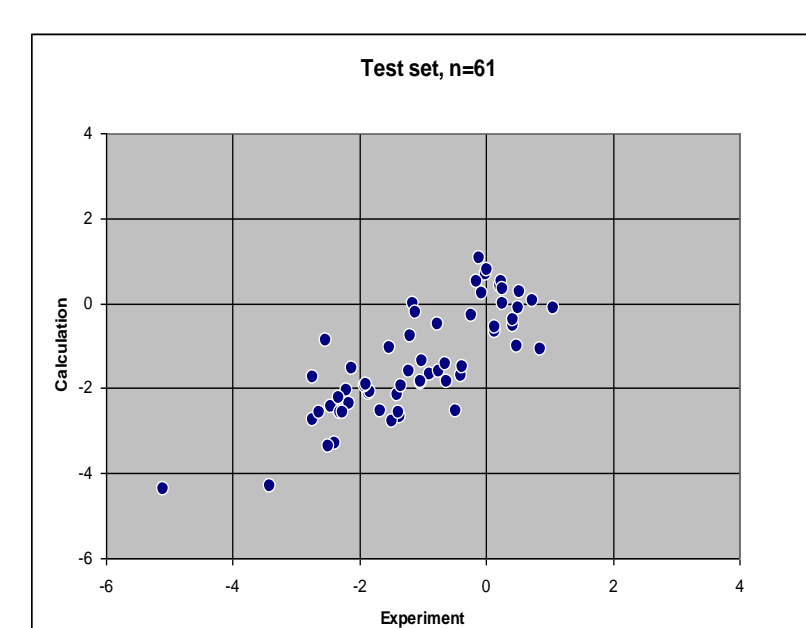
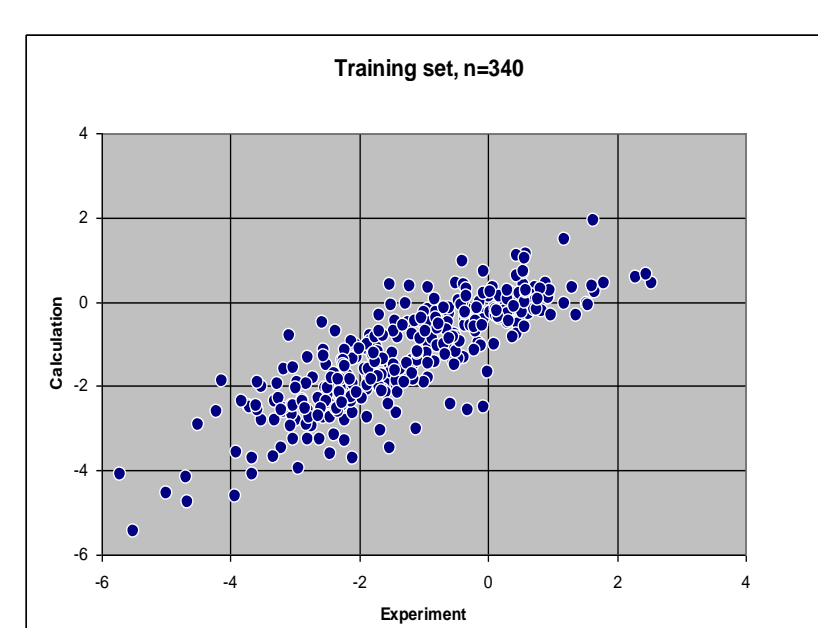


Comparison of the statistical quality of the models for the training (circles) and test (triangles) sets on different values of the LimN. LimN is the minimal value of attributes in the training set which take part in the modeling through an assigned correlation weights.

Example of DCW calculation: SMILES is "O=CC"; CAS is 75-07-0; DCW=1.0018953

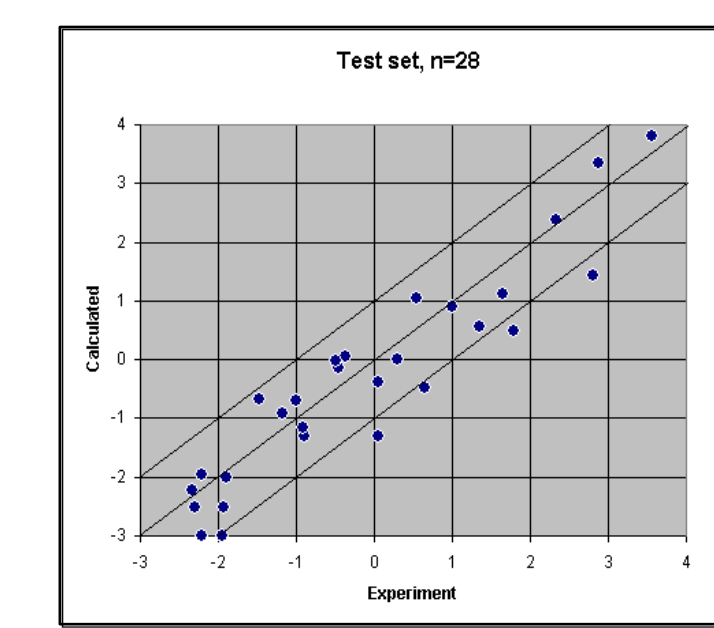
SA _k	CW(SA _k)	N(Train)	N(Test)
O_____	1.0032595	607	121
=_____	0.9918249	471	87
C_____	1.0003857	1547	302
C_____	1.0003857	1547	302
O_=_____	0.9987689	331	67
C_=_____	1.0020759	221	38
C__C_____	0.9969504	401	108
O__=C_____	1.0066095	45	9
C__C__=_____	1.0017028	45	10

Model for carcinogenic potentials



n=340, R2=0.697, s=0.791, F=777 (training set); n=61, R2=0.640, s=0.847, F=105 (test set).

Model for mutagenic potentials



n=67, r2=0.851, s=0.772, F=370 (training set); n=28, r2=0.876, s=0.673, F=183 (test set).

Conclusions

Optimal SMILES-based descriptors are a tool for the QSPR/QSAR analysis

Local and global SMILES attributes can be used in constructing descriptors

Rare SMILES attributes should be blocked since their correlation weights lead to overtraining

Acknowledgements

The authors thank the *Marie Curie Fellowships* for financial support through the contract MIF1-CT-2006-039036 - CHEMPREDICT

