QSAR MODELLING OF MUTAGENICITY: THE APPLICABILITY DOMAIN DEFINITION AND THE ESTIMATION OF PREDICTIVE ABILITY

Andrey A. Toropov[1,2], Alla P. Toropova[1,2] and Emilio Benfenati[2]

[1] Institute of Geology and Geophysics - Khodzhibaev st. 49 - 100041 Tashkent - Uzbekistan
[2] Istituto di Ricerche Farmacologiche Mario Negri - Via Giuseppe La Masa 19 - 20156 Milano - Italy

Introduction

We used simplified molecular input line entry system (SMILES) to construct optimal descriptors for modeling of the mutagenic potency of heteroaromatic amines by quantitative structure - activity relationships (QSAR).
Statistical characteristics of the model are:
n=67, r2=0.8639, r2CV=0.8560, s=0.737, F=413 (training set); and n=28, r2=0.8760, r2CV=0.8560, s=0.644, F=184 (test set)

Materials & Methods

Data on mutagenic potentials of the set of 95 aromatic and heteroaromatic amines was taken from ref. [1]. The mutagenic activity in S. typhimurium TA98 + S9 microsomal preparation is expressed as the natural logarithm of R, where R is the number of revertants per nanomole.
SMILES notations have been generated with the ChemSketch software [2]. The work set (n=95) was randomly split into a training (n=67) and a test set (n=28). Three different splits have been examined (*Figure 1*). These splits are random, but ranges of the mutagenicity are similar for all training sets and all test sets.
The descriptors, which were used in this study, were calculating with SMILES attributes. The SMILES attribute is a combination of SMILES elements. The SMILES element is a group, that contains four, two, or one symbol of a SMILES notation.
In this study 17 SMILES elements have been used: two elements of four symbols are "[N+]" and "[O-]"; three elements of two symbols are "Br", "Cl", and "O="; and 12 elements of one symbol are "(", "1", "2", "3", "4", "C", "F", "N", "O", "S", "c", and "n".
Modeling that is examined in this study included *three steps*:

- *Step 1.* Preparation of list of SMILES attributes for every SMILES notations. Each SMILES attribute is a string of 12 symbols. This string is separated into three zones. The first four symbols is the zone-1; the second four symbols is the zone-2, and the third four symbols is the zone-3. There are three categories of the SMILES attributes. The first category is attributes (1SAk) which are containing sole SMILES element positioned in the zone-1; the second category is attributes (2SAk) which are containing two SMILES elements positioned in zone-1 and zone-2; the third category is attributes (3SAk) which are containing three SMILES elements positioned in zone-1, zone-2, and zone-3. *Table 1* contains an example of the preparation of a list of the attributes for a SMILES notation. In order to avoid a situation when two different SMILES attributes represents the same molecular fragment, for instance the '(N' and the 'N(', the elements for the 2SAk and 3SAk are ranged according to their ASCII codes. Also, the symbol ')' is replaced by '(', because these are representation of the same phenomenon (i.e., branch in molecular skeleton).
- *Step 2.* Preparation of completed list of the SMILES attributes which take place in work set (i.e., in both the training and test sets). Every SMILES attribute is providing by correlation weight equal to 1.
- *Step 3.* Optimization of the correlation weights by the Monte Carlo method. The target function is the correlation coefficient between the log*R* and a descripror, that is calculated with the correlation weights, for the training set. The descriptor is calculated as the following:

$$DCW(limN) = \Sigma CW(1SAk) + \Sigma CW(2SAk) + \Sigma CW(3SAk) \quad (1)$$

where CW(1SAk), CW(2SAk), and CW(3SAk) are the correlation weights for the above mentioned SMILES attributes. The limN is a parameter of the model, that gives possibility to classify the SMILES attributes into two categories: rare or not rare.

Table 1 - Example of the preparation of SMILES attributes: SMILES="Brc1ccc2c3ccc(N)cc3Cc2c1"

No.	^1SA_k	^2SA_k	^3SA_k
1	Br.....		
2	c.....	c...Br.....	
3	1.....	c...1.....	Br..c...1...
4	c.....	c...1.....	c...1...c...
5	c.....	c...c.....	c...c...1...
6	c.....	c...c.....	c...c...c...
7	2.....	c...2.....	c...c...2...
8	c.....	c...2.....	c...2...c...
9	3.....	c...3.....	3...c...2...
10	c.....	c...3.....	c...3...c...
11	c.....	c...c.....	c...c...3...
12	c.....	c...c.....	c...c...c...
13	(.....	c...(.....	c...c...(...
14	N.....	N...(.....	c...(...N...
15	(.....	N...(.....	(...N...(...
16	c.....	c...(.....	c...(...N...
17	c.....	c...c.....	c...c...(...
18	3.....	c...3.....	c...c...3...
19	C.....	C...3.....	c...3...C...
20	c.....	c...C.....	c...C...3...
21	2.....	c...2.....	C...c...2...
22	c.....	c...2.....	c...2...c...
23	1.....	c...1.....	2...c...1...

Results

Our hypothesis is "rare SMILES attributes are able to lead to overtraining".
Influence of the rare attributes may be blocked, if correlation weights of the rare attributes are fixed equal to zero. *Figure 1* demonstrates influence of the limN to statistical characteristics of the QSAR model of the mutagenicity. In detail this approach is described in [3]. *Figure 2* contains an example of the model.

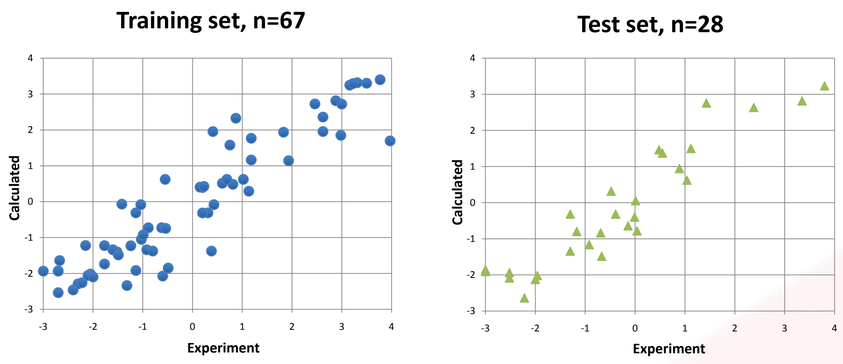

Figure 2 - Example of the QSAR model (limN=6, split 1)

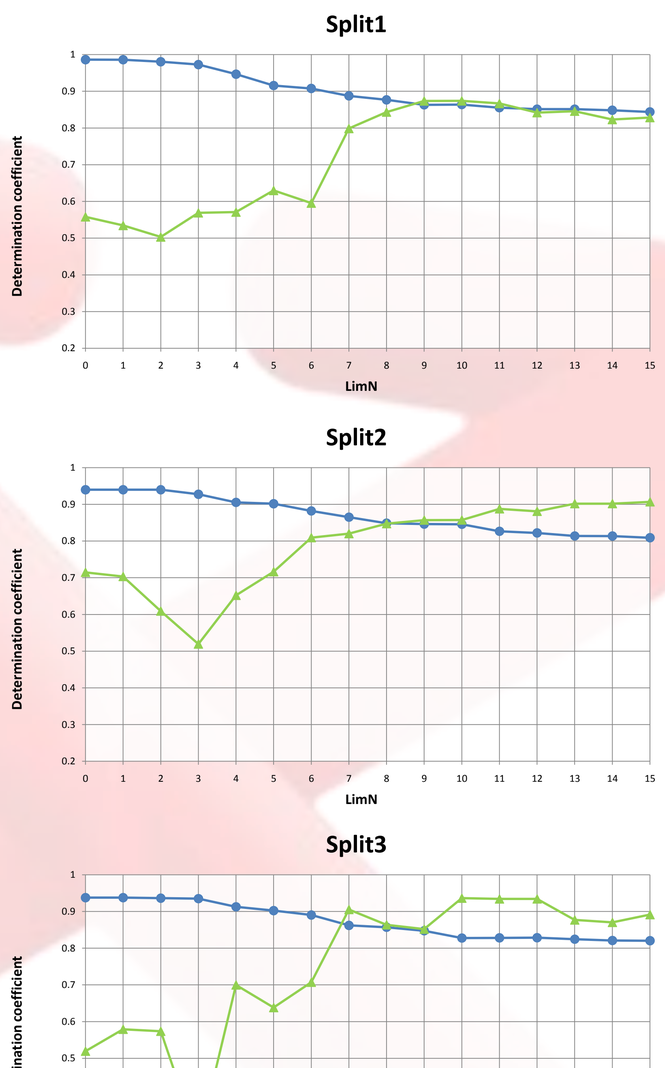

Figure 1 - The influence of blocking of rare SMILES attributes for the training (circles) and the test set (triangles) with limN of 0-15

Conclusions

The one-variables models calculated with the SMILES-based optimal descriptors are satisfactory logR predictors for all three splits into the training and test sets. The blocking of the rare SMILES attributes, by the selecting of the limN-value, improves statistical characteristics of the logR prediction for external test sets (for all three splits). The examined three splits have the different most informative the limN-values (i.e., value that provides the best predicting for the external test set): for the Split1 and Split3 the limN=10, for the Split2 the limN=15. However, it is to be noted that the model for the Split2 with using of LimN=10 is good, though not the best.

References

[1] S. C. Basak, D. Mills, B. D. Gute, R. Natarajan, Top Heterocycl. Chem., 2006, 3, 39
[2] ACD/ChemSketch Freeware, version 11.00, Advanced Chemistry Development, Inc., Toronto, ON, Canada, www.acdlabs.com, 2007
[3] A.A.Toropov, A.P.Toropova, E.Benfenati, Chem. Biol. Drug Des. 2009, 73, 301

Acknowledgements

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