F.T. Adilova [1]
R.R. Davronov [1]
U.U. Zhamilov [1]
Sh.N.Murodov [1]

CASE STUDY: AN INVESTIGATION OF THE KNN-QSAR APPROACH

[1] Institute of Mathematics of the Academy of Sciences of Uzbekistan, Street Do‘rmon yo‘li , 29, Tashkent, Uzbekistan

rifqat@gmail.com

Earlier, we found the dependence of the biological activity of alkaloids of the peganum quinazoline, quinazolone structure and their derivatives with the hydrophobicity of chemical substances on the basis of QSAR models [1]. Criteria for the effectiveness of several types of constructed models were adequacy (in the statistical sense) and interpretability with variations of variables (descriptors) used in the equations. The constructed linear, quadratic, parabolic and bilinear models estimated according to statistical criteria: the determination coefficient R^2, the Fisher test, the value of the confidence interval at the level $p \leq 0.01$. The conclusion from this and our subsequent experiments with QSAR modeling [2] was the need to have an ensemble of models, the best of which in accuracy, and most importantly, in terms of predicted power, should be considered optimal. The aim of this paper is to estimate the efficiency of the kNN-QSAR method [3, 4] using the same experimental data, provided that the calculations are performed by the package R.
On each of the three training samples, using the kNN-QSAR method, 123 different (by descriptor sets) models were constructed, of which 18 were marked as best according to statistical criterions; thus was obtained an ensemble of 18 models. After checking their quality on the test samples, there are 12 models left. In the kNN-QSAR method, it is assumed that the model is suitable for predicting the activity of a chemical compound if $Q^2 > 0.6$. The best predictive models are models with maximum values of Q^2 and F, with a minimum value of RMSE. In Table 1, the best model No. 10 is shown in bold.

Table 1. Statistical evaluation of the quality of the constructed models

Model	q^2	R^2	R_0^2	*RMSE*	*F*	*k*
1	0.81	0.81	0.79	0.003	230.52	1.02
2	0.82	0.95	0.86	0.03	12.83	1.05
3	0.81	0.88	0.65	0.032	42.88	0.94

Thus, the computational experiments showed that the kNN-QSAR algorithm with included k-nearest neighbor procedures, cross-validation and simulation of annealing when selecting descriptors, additional model quality estimates for the two types of samples, generates reliable predictive models of the structure-activity relationships in comparison with simple linear and nonlinear regression equations, taken only from the value of the relative error of the forecast.

1. Adylova F.T. et al. *Kimyovi technology nazorat wa Boscharuve*, 2015, **5:** 21-27 *(in Russian)*
2. Adylova F.T. et al. *Toshkent tibiuet akademiysi akhborotnomasi*, 2016, **2:** *44-48 (in Russsian)*
3. Tropsha A. et al. *Current Pharmaceutical Design*, 2007, **13:** 3494-3504.
4. Tropsha A. *Mol. Inf.*, 2010, **29:** 476-488.