Case Study: Estimating the Predictive Power of the QSAR Model

Fatima, Adilova
Institute of Mathematics
Tashkent, Uzbekistan
e-mail: fatadilova@gmail.com

Rifqat, Davronov
Institute of Mathematics
Tashkent, Uzbekistan
e-mail: rifqat@gmail.com

Uygun, Jamilov
Institute of Mathematics
Tashkent, Uzbekistan
e-mail: jamilovu@gmail.com

ABSTRACT

The intensive development of combinatorial chemistry and high-throughput screening (HTS) in recent years has significantly increased the amount of experimental data of the structure-activity type (SAR). Despite many approaches to the task of estimating, the predictability of QSAR models remains unresolved until now. We preferred to investigate the approach of A.Tropsha and co-authors, who for many years have been developed the kNN-QSAR problem. This paper investigated the problem of strong QSAR model validation on specific data to evaluate the QSAR model predictive capabilities.

Keywords

structure-activity relations, QSAR modelling, model validation.

1. INTRODUCTION

The intensive development of combinatorial chemistry and high-throughput screening (HTS) in recent years has significantly increased the amount of experimental data of the structure-activity type (SAR). This led to the need to use reliable analytical methods, such as the Quantitative Structure-Activity Relation, - QSAR modeling. QSAR is perceived as a tool of establishing correlations between the trends in structural modifications and the corresponding changes in biological activity. However, in many cases, the number of compounds that can be practically synthesized and tested is much less than the total size of virtual chemical libraries. Therefore, the actual need for developing tools for screening virtual libraries is implemented now through QSAR.

The process of developing the QSAR model is divided into three phases: data preparation, model building and validation. Until now, the discussion problem is the validation of the model. Most QSAR modeling methods implement a cross-validation procedure, resulting in a cross-validated R^2 (q^2), used as a criterion of reliability and predictive ability of the model. Many authors consider the high value of q^2 (for example, $q^2 > 0.5$) as an indicator of good predictability of the model. To determine the robustness of the model, Y-randomization (randomization of biological activity) is applied, which consists of repeating the calculation procedure with randomized activities and subsequent estimation of the probability of the resulting statistics (used together with cross-validation, Leave-One-Out, LOO). However, so far the predicted power of the model (even with a high LOO q^2) rarely tested on a set of external data, i.e., compounds that have not been used to develop and internal evaluation of the model.

In spite of many approaches to the task of estimating, the predictability of QSAR models remains unresolved. We preferred the approach of A.Tropsha and co-authors, who for many years have been developed and tested the approach named as kNN-QSAR [1-3]. The aim of this paper is to verify on specific data the QSAR model predictive capabilities in the framework of A. Tropsha's approach.

2. MATERIAL AND METHODS

The data for computational experiments were the results of studies of alkaloids of Peganum harmala L. quinazoline structure and their derivatives[4]. This plant has long been used in folk medicine, but the study of the mechanisms of its pharmacological effect and the creation on this basis of more active and less toxic drugs are still actual problems. From [4], we took 68 compounds that have undergone a complete experimental test, which allows us to estimate the prediction ability of QSAR models.

Let us define quantitative criteria for the predictive ability of the QSAR model. Let $\tilde{y}_i$ and y_i be the predicted and actual activity, accordingly. For an *ideal predictive* QSAR model, the regression line will be the bisector of the angle formed by the positive directions of the orthogonal axes $\tilde{y}$ and y. For an ideal model, the slope of the regression line is 1, the free term is 0, the correlation coefficient R for regression is equal to 1. *The real QSAR model* can have high predictive power if it is close to ideal. This means that the correlation coefficient R between the actual and forecasted activities should be close to 1 and the regression lines $\tilde{y}$ by y or y by $\tilde{y}$ should have slopes k, or k' close to 1.

However, these criteria may not be sufficient for the QSAR model to be predictive. The correlation coefficients for these lines R_0^2 and $R_0^{'2}$ have different values, which are very different from the R^2 values.

Therefore, a more stringent condition is needed that would ensure a high predictive ability: the pairs R^2 and R_0^2, or both $R_0^{'2}$ and R^2 should have the same values. It can be shown that $R^2 \geq \max\{R_0^2, R_0^{'2}\}$. If the angle between the regression lines is small, then these lines are close to each other in the region of their intersection. The proximity R to 1, R^2 to R_0^2 or $R_0^{'2}$ and the corresponding slope k or k' to 1, guarantees the best approximation to the ideal model. The residual root-mean-square error, RMSE and Fisher, F-relation complement the list of statistics. Thus, models are considered acceptable [1], if they satisfy the following conditions: $q^2 > 0.5$, $R^2 > 0.6$, R_0^2 or $R_0^{'2}$ close to R^2, i.e. $[(R^2 - R_0^2)/R^2] < 0.1$,or $[(R^2 - R_0^{'2})/R^2] < 0.1$ and relevant $0.85 \leq k \leq 1.15$ or $0.85 \leq k' \leq 1.15$ (1)

3. RESULTS AND DISCUSSION

All the computational experiments were carried out within the R system. First, using the *rcdk* package, 286 corresponding descriptors were calculated, which were used in constructing the QSAR models. Then all descriptors underwent preliminary filtering, in the process of which descriptors with skips and having pair correlation values greater than 0.7 were excluded. After filtering, the input data matrix consisted of 55 rows (structures) and 65 columns (descriptors). The descriptors were normalized using the *scale* function.

Within the framework of the model validation, the set of initial data is divided into three subsets: training, control and external samples. According to A.Tropsha's recommendations, the number of compounds in the training set should be at least 20, and about 10 compounds should be in the test and external data sets. Considering this, we divided the initial sample three times randomly into a training sample, a test sample, separating the external sample. As a result we received: three training (31 compounds), three test (12 compounds) samples and one external sample of 12 compounds.

As noted above, the robustness of the model is increased by the Y-randomization of activity values. However, until now this issue is debatable in the literature [5,6], so we excluded this procedure from our study. The kNN-QSAR method states: "if an implicit relationship exists between the structure and activity for a given data set, then it can be formally identified using QSAR models obtained with different descriptors and optimization protocols." It follows that several alternative QSAR models should be developed in order to determine the best predictive model for a particular data set. In our case, the procedure for selecting descriptors / variable optimization described above was performed by the methods of simulating annealing (SA) / nearest neighbors for 7 subsets of the original set D of 65 descriptors. Thus, 242, 99, 10 QSAR models were selected according to conditions (1) for the three training sets. Alternative models after processing on test samples were used in a prediction on an external sample. The kNN-QSAR method, which examines all possible combinations of different types of descriptors and optimization methods by final evaluation of the model on an external sample, more accurately takes into account the structure and activity ratio in the original data set. Without dwelling on the details of the method, we note that its result is the QSAR model, which is characterized by a set of informative descriptors identified by the SA procedure and the optimal number of nearest neighbors k.

Three series of computational experiments (CE) were carried out. In each of them, QSAR models were built on training data set; after testing on the test sample, adequate ones were selected, and only they were tested on an external sample to search for a best predictive model. All calculations carried out under the control of the statistical criteria, are described above.

In the CE-1, 242 models were used, in the second computational experiment, -99 models, in the third- CE-3 10 models; all the models had $q^2 > 0.5$ on the training sample. In all three computational experiments, a connection was not found between q^2 and R^2.

Table 1 summarizes the statistics of the best three models predicted in CE-1, CE-2, and CE-3 experiments. From Table 1 it follows that the model obtained in the first series of CE has the greatest predictive power.

Table 1: Values of statistics of the best models

CE	q^2	R^2	R_0^2	RMSE	F
1	**0.72**	**0.92**	**0.87**	**0.005**	**318.88**
2	0.72	0.84	0.83	0.009	82.88
3	0.79	0.78	0.78	0.01	59.19

Figure 1 shows the interdependence between q^2 and R^2 for this model, Figure 2 presents the regression between the observed and predicted activities for compounds from the external data set obtained on the model. As an example, Figure 3 shows the distribution of models by q^2 value, depending on the selected descriptors, and Figure 4 shows a histogram of the number of these models for the same descriptors and q^2.

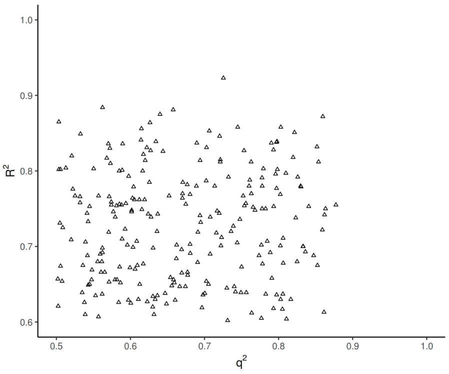

Figure 1: The best forecast QSAR-model. Interrelation q^2 and R^2

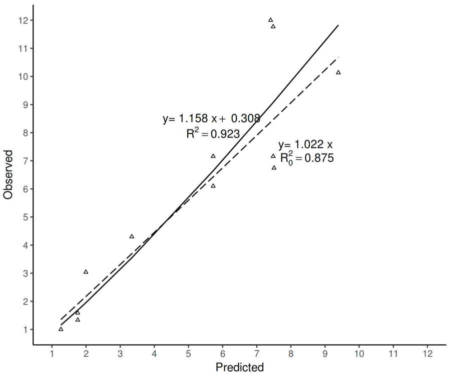

Figure 2: The best forecast QSAR-model. Regression lines between the observed and predicted values

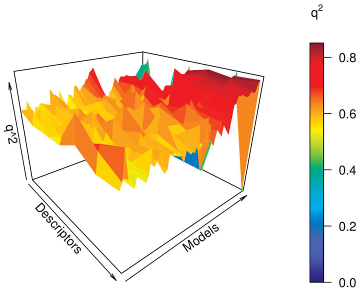

Figure 3: The distribution of models, by q^2 value

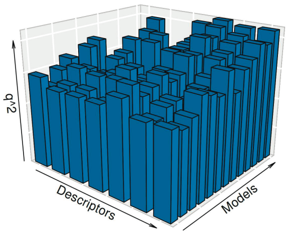

Figure 4: The histogram of the distribution of the number of these models

4. CONCLUSIONS

Previously, we built QSAR models, using many approaches described in the literature for small molecules of various compounds of organic origin[7,8]. However, it has not been possible to obtain a correct proof of their prognostic capabilities so far. The approach of kNN-QSAR, as shown by our computational experiments, is the most adequate, rapid and economical in computational terms. However, questions remain open concerning the applicability domains of the calculated models, and their interpretability, which are the subject of our future research.

5. ACKNOWLEDGEMENT

The research conducted at the Institute of Mathematics under Grant number A-5-1 provided by the Uzbekistan State Committee of Science & Technology.

REFERENCES

[1] A. Golbraikh, A. Tropsha, "Beware of q^2!", *Journal of Molecular Graphics and Modelling*, pp. 269-276, 2002.

[2] A. Tropsha, A. Golbraikh, "Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening", *Current Pharmaceutical Design* , pp. 3494-3504, 2007.

[3] A. Tropsha, "Best Practices for QSAR Model Development, Validation, and Exploitation", *Molecular Informatics*, pp. 476-488, 2010.

[4] N. Tulyaganov, "Pharmacological studies of alkaloids of Peganum Harmala L. quinazoline and quinazolone structure and their derivatives", *The dissertation author's abstract on the scientific degree of Doctor of Chemical Sciences*, Moscow, 1981.(in Russian)

[5] C. Rücker, G. Rücker, M. Meringer, "Y-Randomization A Useful Tool in QSAR Validation, or Folklore?", *http://www.mathe2.uni-bayreuth.de/markus/pdf/pub/YRandQsar.pdf*

[6] C. Rücker, G. Rücker, M. Meringer, "Y-Randomization and Its Variants in QSPR/QSAR", *Journal of Chemical Information and Modelling*, pp. 2345-2357, 2007.

[7] F. T. Adilova, U.U. Jamilov, R.R. Davronov, Sh. N. Murodov, A.A. Azamatov, "LD_{50} prognosis of alkaloids activity in normal quinazoline, quinazolone structure and their derivatives based on QSAR models", *Journal of Chemical control and technology*, pp. 21-27, 2015.(in Russian)

[8] F. T. Adilova, U.U. Jamilov, R.R. Davronov, Sh. N. Murodov, A.A. Azamatov, "Comparative analysis of the development of structure-activity models (QSAR) for a number of diterpene alkaloids: a traditional and a new tools", *Bulletin of the Tashkent Medical Academy*, pp. 44-48, 2016.(in Russian)