A MACHINE LEARNING-BASED QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP STUDY OF A SERIES OF COMPOUNDS FOR THE ANTICANCER ACTIVITY

Sardor Narzullaev [a], Durbek Usmanov [b*], Bakhtiyor Rasulev [a,c]

[a] *National University of Uzbekistan, named after Mirzo Ulugbek, Tashkent, Uzbekistan*
[b] *Institute of the Chemistry of Plant Substances, Academy of Sciences, Tashkent, Uzbekistan*
[c] *Department of Coatings and Polymeric Materials, North Dakota State University, Fargo ND, USA*

Cancer is a collective term used to describe a group of different diseases that are characterized by the loss of control of cell growth and division, leading to a primary tumor that invades and destroys adjacent tissues. It may also spread to other regions of the body through a process known as metastasis, which is the cause of 90% of cancer deaths. Cancer remains one of the most difficult diseases to treat and is responsible for approximately 14.5% of all deaths worldwide. Breast cancer is the most frequent malignancy in females. Due to its major impact on population, this disease represents a critical public health problem that requires further research at the molecular level in order to define its prognosis and specific treatment. Breast cancer is the most frequently diagnosed cancer and the leading cause of cancer death among females, accounting for 23% of the total cancer cases and 14% of the cancer deaths; thus, research in this field is important to overcome both economical and psychological burden.

Application of machine learning in drug development is one of the most powerful modern directions that help to find the best candidates with anticancer activity. One of these methods called cheminformatics that connects machine learning and chemistry. Main cheminformatics approach is Quantitative Structure–activity Relationship (QSAR) methodology that helps to build predictive model of biological activity as a function of structural and molecular information [1-4]. QSAR model is the result of computational process that start with a suitable description of molecular structure and ends with some inference, hypothesis, and predictions on the behavior of molecules in environmental, physicochemical and biological system under analysis. Cheminformatics (QSAR) is widely applied in pharmaceutical industry predictive and diagnostic process which already helped to discover many drugs currently in the market.

This study is devoted to investigation of 105 compounds applying QSAR analysis to correlate and predict their anticancer activity (AA) to the breast cancer cell line (MCF-7). QSAR analysis was carried out using genetic algorithm (GA) for variables selection and multiple linear regression (MLR) analysis. Quantum-chemical descriptors were calculated and applied as well. The model developed shows not only a statistical significance, but also an excellent predictive ability. The estimated predictive ability (r^2_{test}) of the model for the external set is 0.82 and for the training set is $r^2_{train} = 0.89$, respectively.

Materials and Methods

The dataset for the present study has been collected from several experimental studies [5-7] for a series of 105 compounds with anticancer activity (AA) data. All original activity data has been converted into molar 1/log(AA) response variables.

Results and Discussion

The following equation was selected during the study as a best performing one for AA:

$$\begin{gathered}
\mathbf{1/Log(AA)} = 0.001(\pm 0.0005)\mathbf{T(N..F)} + 6.858(\pm 9.022)\mathbf{X2A} + \\
9.937(\pm 3.472)\mathbf{BELm1} + 1.955(\pm 1.510)\mathbf{BELv3} + 0.029(\pm 0.018)\mathbf{RDF080m} + 0.264(\pm 0.211)\mathbf{Mor18u} - \\
0.343(\pm 0.211)\mathbf{Mor21u} - 0.030(\pm 0.097)\mathbf{Mor07m} - 0.155(\pm 0.055)\mathbf{Mor09m} + 21.201(\pm 18.218)\mathbf{G2e} - \\
10.253(\pm 7.711)\mathbf{ISH} - 3.915(\pm 2.455)\mathbf{HATS3m} - 57.537(\pm 31.154)\mathbf{R7u+} - 3.630(\pm 1.665)\mathbf{R1e} - \\
0.081(\pm 0.054)\mathbf{n{=}CR2} - 0.099(\pm 0.126)\mathbf{nCOOR} + 0.087(\pm 0.261)\mathbf{nNHR} - 8.969(\pm 12.499)
\end{gathered}$$

This model shows the best r^2 and q^2 values for the training set, and the best predictive potential for the test set towards AA. Graphical representations of the model are shown in Figure 1.

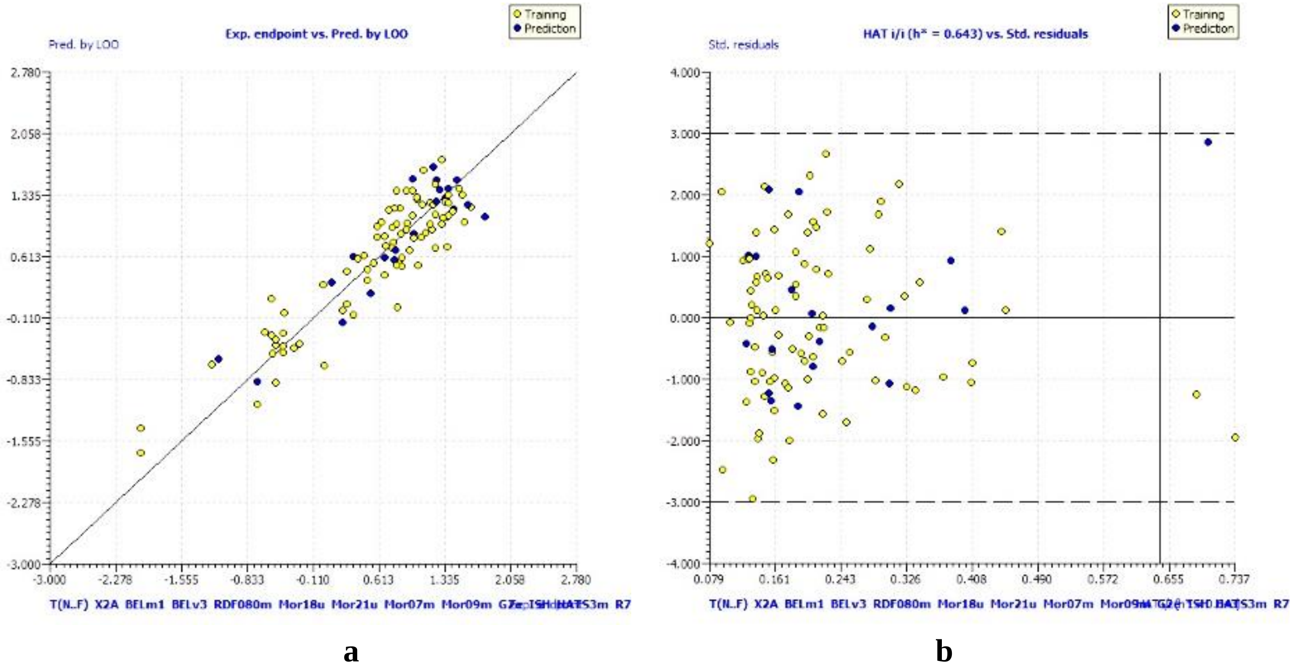

Figure 1. Graphical representation of statistical performance of the best model: (a) Observed vs Predicted correlation plot; (b) Williams plot, all compounds' values are within the applicability domain

The developed machine learning-based QSAR model showed a very good performance and can be now applied to find better candidate compounds with improved anticancer activity. Next steps in this research will be - development of large virtual library of similar class of compounds and screening this library with the developed QSAR model to identify best candidates with high anticancer activity related to the breast cancer.

REFERENCES

[1] Puzyn T., Rasulev B., Gajewicz A., Hu X., Dasari T.P., Michalkova A., Hwang H.M., Toropov A., Leszczynska D., Leszczynski J. Using Nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles, *Nature Nanotechnology*, 2011, 6, 175-178

[2] Turabekova, M.A., Rasulev, B., Dzhakhangirov, F.N., Leszczynska, D., Leszczynski, J., Aconitum and Delphinium alkaloids of curare-like activity. QSAR analysis and molecular docking of alkaloids into AChBP, *European Journal of Medicinal Chemistry,* 2010, 45 (9), 3885-3894

[3] Gajewicz A., Rasulev B., Dinadayalane T., Urbaszek P., Puzyn T., Leszczynska D., Leszczynski J. Advancing risk assessment of engineered nanomaterials: Application of computational approaches, *Advanced Drug Delivery Reviews*, 2012, 64 (15), 1663-1693

[4] Patnode K., Demchuk Z., Johnson S., Voronov A., Rasulev B. Combined Computational Protein-ligand Docking and Experimental Study of Bioplastic Films from Soybean Protein, Zein and Natural Modifiers, *ACS Sustainable Chemistry and Engineering*, 2021, 9, 10740-10748

[5] Abdulrahman, H.L., Uzairu, A. & Uba, S. QSAR, Ligand Based Design and Pharmacokinetic Studies of Parviflorons Derivatives as Anti-Breast Cancer Drug Compounds Against MCF-7 Cell Line. *Chemistry Africa*, 2021, 4, 175–187. doi.org/10.1007/s42250-020-00207-7

[6] Bohari, M. H., Srivastava, H. K., & Sastry, G. N. Analogue-based approaches in anti-cancer compound modelling: the relevance of QSAR models. Organic and Medicinal Chemistry Letters, 2011, 1(1), 3. doi.org/10.1186/2191-2858-1-3

[7] Xu-Yan Wang, Chuang-Jun Li, Jie Ma, Chuan Li, Fang-You Chen, Nan Wang, Cang-Jie Shen, Dong-Ming Zhang. Cytotoxic 9,19-cycloartane type triterpenoid glycosides from the roots of Actaea Dahurica. *Phytochemistry*, 2019, 160, 48-55, doi.org/10.1016/j.phytochem.2019.01.004.