Extension of pQSAR: ensemble model generated by Random Forest and Partial Least Squares Regressions

Byung Chun Kim [1] , Dosang Joe [1] , Youngho Woo [1] , Yongkuk Kim [2] and Gangjoon Yoon [1]

1) *National Institute for Mathematical Science, 70, Yuseong-daero 1689 beon-gil, Yuseong-gu, Daejeon 34047 Korea*

2) *Kyungpook National University, 80 Daehakro, Bukgu, Daegu 41566 Korea*

Speaker : Byung Chun Kim, bckim@nims.re.kr

ABSTRACT

Quantitative structure-activity relationship (QSAR) regression models are mathematical ones which relate the structural properties of chemicals to the potencies of the biological activities of the chemicals. In QSAR models, the physical and chemical information of the molecules is encoded into quantitative numbers called descriptors. Recently, experimental test results (profiles) have been used as descriptors of chemicals. Profile QSAR 2.0 (pQSAR) model suggested by Martin et. al, is a multitask, two step machine learning prediction method with a combination of random forest regressions (RFRs) and partial least squares regression (PLSR). In pQSAR model, one fills the profile table's missing values with RFRs and then builds PLSR using the profile predictions. Note that in the second step of the pQSAR method, PLSR's predictor variables are profiles; so activity values, and the response variables are also activity values. Thus we can use the PLSRs to update the profile table and then repeat the second step. In this work, we propose an extended model of pQSAR generated by RFRs and PLSRs. Experiment of updating the given full initially predicted profile table by two kinds of prediction models, RFRs and PLSRs, has been conducted iteratively for the PKIS and ChEMBL data sets. Even though prediction performance of individual combination of RFRs and PLSRs varies, the average of the all possible predicted profile tables for given iteration shows better performance. This ensemble model has better prediction performance in sense of Pearson's R^2 compared to that of the pQSAR model.

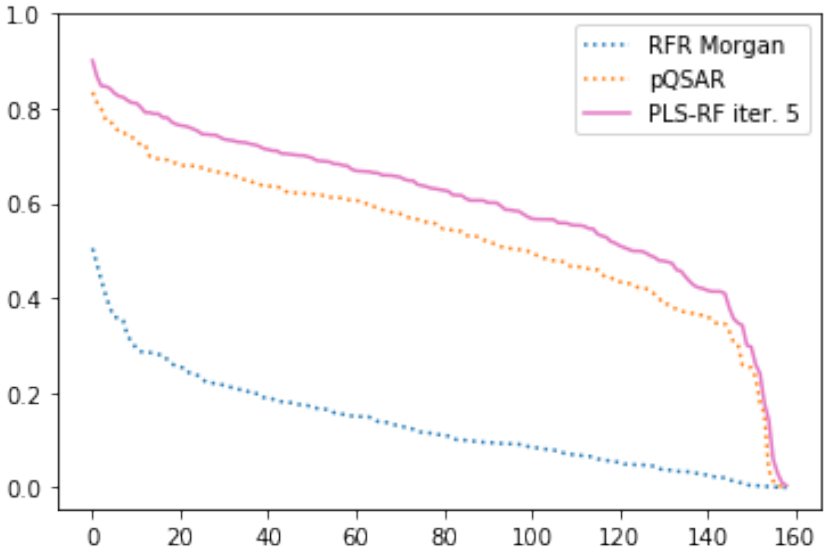

Figure 1. R^2 distribution for the ChEMBL data. The graph was plotted in descending order of R^2 values for the three models.

Table 1 R^2 performance comparison for the ChEMBL assay data.

R^2	ave.	median	≥ 0.3	≥ 0.5	≥ 0.6	≥ 0.7
RF (M)	0.13462	0.11148	10	1	0	0
pQSAR	0.5269	0.55026	148	99	63	13
PLS-RF						
iter. 1	0.58411	0.61225	149	122	87	39
iter. 2	0.59104	0.61787	149	119	88	42
iter. 3	0.59925	0.62822	**151**	121	88	**49**
iter. 4	0.59808	0.62380	149	121	91	48
iter. 5	**0.60058**	**0.62868**	150	**123**	**93**	49

REFERENCES

1. E. J. Martin and V. R. Polyakov and X. Zhu and L. Tian and P. Mukherjee and X. Liu, "All-Assay-Max2 pQSAR: Activity Predictions as Accurate as Four-Concentration IC50s for 8558 Novartis Assays", *Journal of Chemical Information and Modeling*, Vol. 59 (10), 2019, pp. 4450-4459.
2. A. P. A. Janssen, S. H. Grimm, R. H. M. Wijdeven, E. B. Lenselink, J. Neefjes, C. A. A. van Boeckel, G. J. P. van Westen, and M. van der Stelt, "Drug Discovery Maps, a Machine Learning Model That Visualizes and Predicts Kinome–Inhibitor Interaction Landscapes", *Journal of Chemical Information and Modeling*, Vol. 59 (3), 2019, pp. 1221-1229