On a New Method of Reducing the Set of Compounds to Improve QSAR Models Using Molecular Signatures

Fatima Adilova
Institute of Mathematics
Tashkent, Uzbekistan
e-mail: fatima_adilova@rambler.ru

Alisher Ikramov
National University of Uzbekistan
Tashkent, Uzbekistan
e-mail: ikramov.alisher@list.ru

ABSTRACT
This paper provides the description and analysis of a new method of reducing the general sample of compounds using molecular signatures developed by J.-L. Faulon. This method shows the improvement of the resulting QSAR models.

Keywords
QSAR, molecular signatures, set reduction, machine learning.

1. INTRODUCTION
Molecular modeling is an important research tool in many fields of chemistry. Molecular signatures were developed by J.-L. Faulon [1-3] for enumerating isomers. Then it was used to enumerate molecules [4,6] and to reconstruct molecules matching molecular descriptors to solve the inverse QSAR problem [5].

The problem of set reduction is well analyzed. Cross- validation [7,10] involves partitioning a data sample into two subsets, performing the analysis on one sub- set (training set), and running the analysis on the other subset (testing set). Multiple rounds of cross-validation are performed using different partitions on two subsets. Using this approach one can find a better partitioning and exclude those compounds from the test set that are badly analyzed.

Bootstrap [8] constructs resamples that are the same in size but contain some compounds not once. Re- samples are used as training sets and the initial sample as a test set. Those resamples that have better results can form the finite set (after removing the repeated com- pounds).

Both approaches were compared in [9]. Our new approach does not require an activity value. This is the main difference from the other mentioned methods. The main goal is to get a sample that complies with QSAR paradigm — similar compounds have similar activity [11]. The sample reduction based on removing all compounds that have a rare or unique structure helps to reach this goal.

2. MATERIALS AND METHODS
We chose 1794 compounds from ChEMBL Data Base with their activity values against human carbonic anhydrase II and descriptors provided by ChEMBL. Then we computed their molecular signatures using MolSig Software[1]. We combined both descriptors in one data set and run our program that looked for all signatures that appear only in less than T compounds. The program removed all such signatures and all compounds that contain these signatures. We obtained 742 different signatures overall.

After the reduction, we randomly permutated the finite sample and divided it onto subsets with 200 compounds in each. Then we generated two files for each subset — one with an activity class and ChEMBL descriptors only, the other with an activity class and molecular signatures only.

3. RESULTS AND DISCUSSION
SVM models were trained on each generated file and tested on other files (with the same type of descriptors and the same T).

As number T is not fixed it is important to observe changes in effectiveness of SVM models while increasing T. We used three different values — 5, 9, and 13. The comparative results are presented in Table 1, where we expressed only the average values of efficiency (efficiency of model is counted by SVM-predict Software using the trained model and test set). It shows the increase in predictive efficiency of SVM models trained not only on molecular signatures but on ChEMBL descriptors that were not used during the reduction process.

Table 1: Comparison of SVM models' average efficiency by changing threshold in reduction algorithm

T-value	ChEMBL	Signatures
5	44.7%	55.0%
9	50.5%	61.2%
13	59.1%	65.0%

The reduction algorithm produced sizes and number of signatures depending on T-value presented in Table 2. All signatures after reduction appear in at least T com- pounds each in resulted samples. As the sample for T = 13 can generate only three subsets sized 200, we run random permutation twice and produced two in- dependent families of subsets. Each model trained on a subset from one family was tested on all subsets from only the same family.

[1]https://sourceforge.net/projects/molsig/

We also examined the efficiency of SVM models testing on training sets and got average 60 % on ChEMBL descriptors and 99 % on signatures. We did not include these results with consideration as the efficiency of SVM on signatures on the same set can be explained by a high dimension value. So it would be inaccurate to use these results in Table 1.

Table 2: Size of data sets and number of signatures generated by reduction algorithm

T-value	Size	Number of Signatures
5	1358	342
9	1004	246
13	643	151

4. CONCLUSIONS

Removing process of rare signatures along with the correspondent compounds showed an increase of efficiency of not only the models trained on molecular signatures but also on ChEMBL descriptors. This approach allows to reduce the set of compounds without using their activity value. The research examined different threshold values for the reduction algorithm and proved its importance.

Although the dimension of signatures space is much larger than the number of ChEMBL descriptors (which is 19), the SVM models trained on signatures are quite more effective than those trained on ChEMBL descriptors.

5. ACKNOWLEDGEMENT

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

REFERENCES

[1] J.-L. Faulon. "On Using Graph-Equivalent Classes for the Structure Elucidation of Large Molecules". *J. Chem. Inf. Comput. Sci.*, 32, pp. 338-348, 1992.

[2] J.-L. Faulon. "Stochastic Generator of Chemical Structure. 1. Application to the Structure Elucidation of Large Molecules". *J. Chem. Inf. Comput. Sci.*, 34, pp. 1204-1218, 1994.

[3] J.-L. Faulon. "Stochastic Generator of Chemical Structure. 2. Using Simulated Annealing To Search the Space of Constitutional Isomers". *J. Chem. Inf. Comput. Sci.*, 36, pp. 731-740, 1996.

[4] J.-L. Faulon, C.J. Churchwell. "The Signature Molecular Descriptor. 2. Enumerating Molecules from Their Extended Valence Sequences". *J. Chem. Inf. Comput. Sci.*, 43, pp. 721-734, 2003.

[5] W. M. Brown, S. Martin, Mark D. Rintoul, J.-L. Faulon. "Designing Novel Polymers with Targeted Properties using the Signature Molecular Descriptor". *Journal of Chemical Information and Modeling*, 46(2), pp.826-835, 2006.

[6] P. Carbonell, L. Carlsson, J.-L. Faulon. Stereo signature molecular descriptor. *Journal of Chemical Information and Modeling*, 53 (4), pp. 887897, 2013. doi: 10.1021/ci300584r

[7] Y. Bengio, Y. Grandvalet. "Bias in estimating the variance of k-fold cross-validation". *Statistical modeling and analysis for complex data problems*, pp.7595, 2005.

[8] W. Jiang, R. Simon. "A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification". *Statistics in Medicine*, 26(29), pp.53205334, 2007.

[9] J.-H. Kim. "Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap". *Computational Statistics and Data Analysis*, 53(11), pp.37353745, 2009. doi:10.1016/j.csda.2009.04.009

[10] RJ. Tibshirani, R. Tibshirani. "A bias correction for the minimum error rate in cross-validation". *Arxiv preprint* arXiv:0908.2904, 2009.

[11] A.K. Debnath. Quantitative structure-activity relationship (QSAR) paradigm–Hansch era to new millennium. *Mini Rev Med Chem*, 1(2), pp.187-95, 2001 Jul.