F.T.Adilova [1]
A.A.Ikramov [2]

AN APPROACH OF REDUCING COMPOUNDS' SET BASED ON MOLECULAR SIGNATURES

[1] Laboratory of Bioinformatics, Institute of Mathematics, Uzbekistan;
[2] Faculty of Mathematics, National University of Uzbekistan;
fatima_adilova@rambler.ru, melan44@mail.ru

Extensive research shows that it is often impossible to build QSAR models with good predictive power, even when using the most sophisticated algorithms and meticulous simulation. Often there are compounds classes that are presented insufficiently in sample. Therefore models built on such samples are not effective on compounds from these classes, and moreover are not effective on other compounds from the sample.

There are many approaches to clean out the samples from such compounds. *The widespread adoption of HTS and combinatorial chemistry techniques led to a surge of interest in chemical diversity. It was widely expected that simply making "diverse" libraries would provide an increase in the number of "hits" in biological assays. However, it was soon realised that merely making large number of molecules was not sufficient; it was also important to take other properties into account. There are[1] four main approaches to select diverse sets of compounds: cluster analysis, dissimilarity-based methods, cell-based methods and the use of optimisation techniques.*

Therefore, the selection of diverse subsets is based on the premise that structural similarity is related to similarity in biological activity space. Brown and Martin [2] in their clustering experiments concluded that compounds within 0.85 Tanimoto similarity (calculated using UNITY fingerprints) have an 80% chance of sharing the same activity. However, recent work by Martin et al. [3] shows that the relationship between structural similarity and biological activity similarity may not be so strong as these previous studies suggested. They found that a molecule within 0.85 similarity of an active compound had only a 30% chance of also being active.

We here present another approach that showed good performance independent from descriptors space.

We studied signature descriptors approach from [4,5] to build efficient QSAR models. We used 1794 compounds from ChEMBL with activity value against human carbonic anhydrase II. The sample was divided randomly on several sets with 200 compounds each. Each set was stored in two files – one with ChEMBL descriptors, and other with signatures (they were calculated using Software https://sourceforge.net/projects/molsig/). We run SVM to develop models on each set and tested on other sets. We got average 43.7 % efficiency on ChEMBL descriptors and average 53.3 % efficiency on signatures. But we had 794 different signatures generated. Some of signatures appeared only in 1 or 2 compounds. So, we decided to cut all compounds that have rare signatures (that appear within 5 or less compounds only) and thus those signatures. The sample became of 1358 compounds with only 342 signatures. Each molecular signature appears in at least 6 compounds.

We run the same process of dividing the sample onto smaller sets and training SVM on them. We got average 50 % efficiency on ChEMBL descriptors and average 63 % efficiency on signatures.

Therefore, we consider the signature based set reducing approach as very perspective even if researchers do not use them in models they develop.

1. Leach A.R. et al. Introduction to Chemoinformatics, Second Edition, Springer, 2007.
2. Brown R.D. et al. *Journal of Chemical Information and Computer Sciences*, 1996, **36**: 572–583.
3. Martin Y.C. et al. *Journal of Medicinal Chemistry*, 2002, **45**: 4350–4358.
4. Brown W.M. et al. *Journal of Chemical Information and Modeling*, 2006, **46(2)**: 826-835.
5. Carbonell P. et al. *J Chem Info Model*, 2013, doi: 10.1021/ci300584r