[SC6] HTS-likeness: physicochemical parameters to create libraries

Pavel Polishchuk, Mariia Matveieva

Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University Olomouc, Hněvotínská 1333/5, 779 00 Olomouc, Czech Republic

Creation of compound libraries for high-throughput screening (HTS) is one of the strategies to discover new active compounds. It becomes more and more affordable for academia to create own HTS libraries. However, due to limited budgets those libraries should be relatively small and at the same time provide high chance to discover true actives. Proper selection of compounds is a hard task due to different issues. Selected compounds should satisfy certain physicochemical criteria (drug-likeness or lead-likeness), they should not be reactive, toxic, etc, and they should be diverse. In our study we investigated the influence of different physicochemical parameters in order to find better criteria for compound selection that existing lead- or drug-likeness rules.

We collected data about MLSMR compound library which was tested in numerous assays and which data are available on PubChem. This library was created taking into account drug-likeness. Therefore, it might be challenging to find better selection criteria for compounds based on this set. We split data on train and test sets. The train set contained 230325 compounds which all were tested in 49 assays. The test set contained 72670 tested in 44 assays. We considered cell-based and biochemical assays separately because compounds active in these assays might have different physicochemical properties. We analyzed distributions of different physicochemical properties of train set compounds and how these influence the hit rate across assays. It was established that compounds active in different types of assays really have different profiles. Manually derived rules demonstrated some advantage over common drug-likeness rules: the number of selected compounds was decreased substantially and the hit rates were increased in many assays. However, manual rule derivation might be not optimal. Therefore, we implemented special decision tree and random forest algorithms and used them to build models which predict HTS-likeness of compounds and can be used for compound selection.

The decision tree model provided only slight improvement for compound selection over manually derived rules. Rules derived from decision tree models were somewhat different from manually derived ones and selected the less number of compounds. Random Forest models are not easily interpretable but provided clear advantage over manually derived and decision tree rules. These models selected the reasonably small number of compounds from the test set and substantially improved chances to find actives. These models can be recommended for creation of HTS libraries.

Acknowledgement:
This work is supported by the grant of the Ministry of education, youth and sport of Czech Republic (MSMT-5727/2018-2).