Composition of the set of inactives and the performance in the classification of bioactive compounds by machine learning methods

Sabina Smusz, Rafał Kurczab, Andrzej J. Bojarski

Institute of Pharmacology, Polish Academy of Sciences, Kraków, Poland

Machine learning methods are among the most popular tools used in cheminformatics and its major task is the assignment of objects (here: molecules) into classes (active or inactive). Many different algorithms are used in this methodology: decision trees, support vector machines and Naïve Bayes classifier are just a few examples of numerous approaches to deal with the classification problems.[1] It has been already proved, that machine learning methods performance strongly depend on various factors, such as the type of fingerprint used for molecules representation, number of active compounds in the training data and application of meta-learning.[2]

Here, we present a study on the influence of the set of inactives composition on the classification effectiveness. Six different ways of selecting compounds inactive (or assumed as inactive) towards 5 protein targets were tested: random selection (10 different sets for each target) from the ZINC database, MDDR database and libraries generated according to the DUD methodology[3], as well as selection from already mentioned compound libraries, with the assurance of maximum diversity of the chosen structures.

This study is supported by project UDA-POIG.01.03.01-12-100/08-00 co-financed by European Union from the European Fund of Regional Development (EFRD).

[1] Melville, J. L.; Burke, E. K.; Hirst, J. D. *Comb. Chem. High Throughput Screen.* **2009**, *12*, 332-343.
[2] Kurczab, R.; Smusz, S.; Bojarski A. J. *J. Cheminform.* **2011**, *3*, P41.
[3] Huang, N.; Shoichet, B. K.; Irwin, J. J. *J. Med. Chem.* **2006**, *49*, 6789-6801.