Active or inactive: that is the question. An application of machine learning methods in virtual screening

Sabina Smusz

Virtual screening of large databases of compounds is widely used in drug discovery campaigns. The search for new ligands may be based on the information about known actives (ligand-based VS) or it may rely on the knowledge about a protein target (structure-based VS).

In recent years, an application of machine learning methods in virtual screening is becoming increasingly popular. In this case, the task of machine learning algorithm is to classify molecules as active or inactive ones. We have tested different supervised machine learning approaches, such as: support vector machines, decision trees, lazy classifiers, meta classifiers and many others using tools implemented in WEKA package. Classifiers were trained and tested on sets containing compounds from our internal database of known 5-HT7 antagonists. We have examined their performance depending on the number of active compounds present in the training set, the type of fingerprints and the type of filter used for attributes selection in data preprocessing.