

Is There a General Model for Bioactivity?

T.I. Oprea, O. Ursu, C.G. Bologa, and L.A. Sklar
New Mexico Molecular Libraries Screening Center
1 University of New Mexico, MSC08-4630
Albuquerque, NM 87131-0001, USA

The Molecular Libraries Screening Centers Network (MLSCN) uploads bioactivity screening data in PubChem (<http://pubchem.ncbi.nlm.nih.gov/>) based on the Molecular Libraries Small Molecules Repository (MLSMR). On 01-30-2008, we found 222,878 unique MLSMR compounds tested on 478 MLSCN bioassays. Not all compounds were tested on all assays.

The majority of MLSMR compounds are “inactive” in the above numbers of assays. A smaller subset of MLSMR were active (25.03%, active in >1 assay), or among inconclusives (30.56%, in > 1 assay). Overall, a large percentage of compounds are inactive in all assays to date. This dataset allows us to address the following question: Is there a general model for bioactivity?

We used chemical fingerprints (560 predefined keys) and a machine learning technique (support vector machines, SVM), to discriminate actives from inactives. For “inactives” we used compounds that were found inactive in at least 60 assays; all actives or inconclusives were removed from this dataset, to yield over 22,271 compounds. For “actives” we used compounds that were found active in at least one assay; all “inactives” were removed from this data set to yield over 55,782 compounds. To the “actives” set of compounds were added the compounds from WOMBAT 2007.1 database (<http://www.sunsetmolecular.com>) to yield over 209,451 unique compounds.

From a pool of 209,451 actives and 22,271 inactives, two sets of 6000 randomly selected compounds were used to build/validate 100 different active/inactive models using a Radial Basis Function SVM kernel. Each SVM model was build using 300 random “actives”/“inactives” compounds from 6000 subset and was validated using the rest of the subset.

External prediction yields ~67% accuracy in the active class (135,467 out of 203,307 actives), and ~83% for the inactive class (13,350 out of 16,127 compounds).

An analysis based on molecular weight shows a small shift towards higher molecular weight for the “actives” set compared to the “inactives” set. Analyses based on estimated aqueous solubility and octanol/water partitioning did not indicate significant differences between “actives” and “inactives”. Furthermore, a large number of chemical scaffolds are present in both the active and inactive class. Artfactual results, such as florescent compounds, and aggregators, were not individually examined. However, in one MLSMR/MLSCN assay, only ~1100 compounds out of ~70,000 were considered potential aggregators.

Taken together, these results appear to indicate that bioactivity, as captured in the MLSMR and WOMBAT “actives”, can be discriminated from the MLSMR “inactives”. If validated by additional data, such models could be used to enrich screening libraries with compounds that are more likely to belong to the “active” class.