[P9] Generative Topographic Mapping in Virtual Screening: why ensemble of maps is needed?

Yuliana Zabolotna[1], Iuri Casciuc[1], Dragos Horvath[1], Gilles Marcou[1], Jürgen Bajorath[2], Alexandre Varnek[1]

[1] *Laboratory of Chemoinformatics, University of Strasbourg, France*
[2] *B-IT, Limes, Unit Chem. Biol. & Med. Chem., University of Bonn, Germany*

Generative Topographic Mapping (GTM)[1] has already been proven as a versatile tool in QSAR modeling[2]. Here, we report an application of a new "universal GTM" approach to virtual screening (VS). A universal GTM[3] represents a map selected for its "polypharmacological competence", *i.e.* its ability to simultaneously host meaningful activity and property landscapes, associated to many distinct targets and properties. However, several such GTMs, only slightly differing with respect to their ability to separate "actives" from "inactives" on associated target classification landscapes, can be generated – each based on a different initial descriptor vector, encoding distinct structural features. While their average "polypharmacological competence" may indeed be similar, they may nevertheless significantly diverge with respect to the quality of each property-specific landscape. In this work we have shown that distinct "universal" maps may represent complementary points of view on chemical space – each based on different descriptors capturing distinct structural aspects, allowing them to provide better landscapes for certain properties that are well explained by underlying descriptors.

Eight such distinct "universal" GTMs were employed as support for predictive classification landscapes, for more than 600 ligand series associated to as many ChEMBL23 targets. For nine of these targets, it was possible to extract, from the Directory of Useful Decoys (DUD)[4], truly external sets featuring sufficient "actives" and "decoys" not present in the landscape-defining ChEMBL ligand sets. For each such molecule, projected on every class landscape of particular universal map, a probability to be active or inactive was estimated.

For a given activity type, no correlation between Balanced Accuracy (BA) values obtained for 8 universal maps in cross-validation on ChEMBL data and in external validation on DUD data was observed. Thus, it would be an error to prefer the universal map with best cross-validation results for a given property as the implicitly best predictor. Moreover, depending on the used map, predictions for many DUD compounds are deemed as not trustworthy, according to applicability domain considerations. By contrast, simultaneous application of all 8 universal maps, and rating of the likelihood to be active as the mean likelihood returned by all applicable maps, significantly improved both the prediction results and the rate of trustworthy predictions. Thus, for a given target, consensus BA was similar to that for the best individual map. Almost 100% of external compounds could be predicted in this way (they were present within the applicability domain of at least one of the eight maps). The consensus value of Enrichment Factor (EF) calculated for top 100 compounds was always significantly larger than EF obtained with any individual map. Thus, the distinct "universal" GTMs are indeed highly complementary and synergistic.

Bibliography:

1. Bishop, C. M.; Svensén, M.; Williams, C. K., *Neural computation* 1998, *10* (1), 215-234.
2. Gaspar HA, Baskin II, Marcou G, Horvath D, Varnek A..*J.ChemInf.Mod.* 2015, 55(1), 84-94.
3. Sidorov P, Gaspar H, Marcou G, Varnek A, Horvath D. *JJ. Comp.Aided Mol.Des.* 2015, 29(12), 1087-108.
4. Huang N., Shoichet B. K., Irwin J. J. *J. Med. Chem.*, 2006, 49 (23), pp 6789–680d1