LIT-PCBA: An unbiased database for evaluation of virtual screening methods

Tran-Nguyen V.K.,[1] Bret G.,[1] Rognan D.[1,*]

[1] *Laboratoire d'Innovation Thérapeutique, Université de Strasbourg, 74 Route du Rhin, 67400 Illkirch-Graffenstaden, France*

* Correspondence: rognan@unistra.fr

Serious biases related to commonly used datasets for retrospective structure-based virtual screening studies (e.g. DUD, DUD-E, ChEMBL) have been reported in recent years [1,2]. The composition of each dataset has been heavily biased, as the quantity of active compounds is usually too high; the potency of presumably inactive compounds always remains unknown; and the actives are too similar (in 2D) to each other and to the crystallographic reference structures deposited on Protein Data Bank, while remarkably different from the inactive counterparts. Such datasets do not mimic chemolibraries used for high throughput screening in reality, tend to overestimate the performance of virtual screening methods, and are not recommended for benchmarking purposes. The need to design a novel and unbiased database dedicated to structure-based *in silico* screening approaches therefore arises. We herewith present the newly designed LIT-PCBA database, consisting of 21 datasets representing 11 protein families of pharmaceutical interest (including kinases, GPCRs, nuclear receptors and other targets), which was constructed based on the experimental results of biological tests deposited on PubChem's BioAssays, thus confirming the potency of active and inactive compounds [3]. All substances were prepared and filtered in such a way that assay artifacts (false positives) as well as artificial enrichment were prevented according to our selection rules and those previously explained by Rohrer S.G. and Baumann K. [4]. The ratio between the number of active compounds and that of inactives has been greatly reduced, and the potency of remaining actives is remarkably lower than that found in the DUD-E database. Retrospective virtual screening results using two ligand-based methods (2D geometry similarity search by ECFP4 and 3D geometry similarity search with ROCS) and a structure-based approach (molecular docking with Surflex-Dock) show that screening performances (ROC AUC, BEDROC AUC, EF1%) varied depending on the PDB template structure that was used for each set and the method that was employed, and there is little structural bias that remains among the compounds that constitute most datasets. The LIT-PCBA database can therefore be used to compare the real accuracy of scoring functions in future benchmarking research.

Bibliography :

[1] Chaput L. et al. J. Cheminform., 2016, 8, 56.
[2] Sieg J. et al. J. Chem. Inf. Model., 2019, 59, 947-961.
[3] https://pubchem.ncbi.nlm.nih.gov/ (accessed Jul 31, 2019).
[4] Rohrer S.G.; Baumann K.J. Chem. Inf. Model., 2009, 49, 169-184.