[P45] Computational chemogenomics - is it more than inductive transfer?

J.B. Brown[1], Yasushi Okuno[1], Gilles Marcou[2], Alexandre Varnek[2], Dragos Horvath[2]

[1] *Kyoto University Graduate School of Pharmaceutical Sciences, Department of Systems Bioscience for Drug Discovery, 606-8501, Kyoto, Japan*
[2] *Laboratoire de Chémoinformatique UMR 7140 CNRS/UdS, Université de Strasbourg, 1 rue B. Pascal, 67037, Strasbourg, France*

High-throughput assays challenge us to extract knowledge from multi- ligand, multi-target activity data. In QSAR, weights are statically fitted to each ligand descriptor with respect to a single endpoint or target. However, computational chemogenomics (CG) has demonstrated benefits of learning from entire grids of data at once, rather than building target-specific QSARs. A possible reason for this is the emergence of inductive knowledge transfer (IT) between targets, pro- viding statistical robustness to the model, with no assumption about the structure of the targets. Relevant protein descriptors in CG might allow to learn how to dynamically adjust ligand attribute weights with respect to protein structure. Hence, models built through explicit learning by including protein information (EL), while benefitting from IT enhancement, should provide additional predictive capability, notably for protein deorphanization.

This interplay between IT and EL in CG modeling is not sufficiently studied. While IT is likely to occur irrespective of the injected target information, it is not clear whether and when boosting due to EL may occur. EL is only possible if protein description is appropriate to the target set under investigation. The key issue here is the search for evidence of genuine EL exceeding expectations based on pure IT.

We explore the problem in the context of Support Vector Regression, using >9400 pK_i values of 31 GPCRs, where compound-protein interactions are represented by the concatenation of vectorial descriptions of compounds and proteins. This provides a unified framework to generate both IT-enhanced and potentially EL-enabled models, where the difference is toggled by supplied protein information. For EL-enabled models, protein information includes genuine protein descriptors such as sequence counts. EL- and IT-based methods were benchmarked alongside classical QSAR, with respect to cross-validation and deorphanization challenges.

While EL-enabled strategies outperform classical QSARs and favorably compare to similar published results, they are, in all respects evaluated, *not* strongly distinguished from IT-enhanced models. Moreover, EL-enabled strategies failed to prove superior in deorphanization challenges.

Therefore, this paper argues that, contrarily to common belief and intuitive expectation, the benefits of chemogenomics models over classical QSAR are actually less due to the injection of protein-related information, but rather the effect of inductive transfer, due to simultaneous learning from all the modeled endpoints. These results show that the field of protein descriptor research needs further improvements to realize the expected benefit of EL.