Comprehensive Variable Selection Strategies in Linear & Non-Linear QSAR

[a]Dragos Horvath, [b]Alexandre Varnek.

[a] UMR8525 - CNRS
Institut de Biologie de Lille
Campus Institut Pasteur 1, rue Calmette 59027 Lille Cedex France

[b]Faculté de Chimie, Université Louis Pasteur,
4, rue B. Pascal, Strasbourg 67000, France

The problem of selecting the appropriate descriptors to enter a QSAR model has been traditionally addressed in an almost artisanal fashion, by practical limitations in terms of descriptor sets one was actually able to generate, trials and error, redundant descriptor dismissal (but with no agreement on how redundant they need to be...). However, robust reproducible and automated selection strategies are needed in order to maximize the reliability of QSAR approaches. Automated descriptor selection procedures are either deterministic (such as stepwise regression - which only worksfor linear models, and does not ensure that the overall optimal selection will effectively be reached by the "greedy" strategy of entering the locally most discriminant term into the equation) or stochastic - but then, the computer effort allocated to solve the problem was not always adapted to the huge phase space (possible descriptor combinations) to sample.

We took profit of the recent development of topological fuzzy pharmacophore triplets, a molecular fingerprint of typical dimension (e.g. number of components) of several thousands, to explore novel variable selection schemes - both stochastic and deterministic. The stochastic approach relies on a parallellized genetic algorithm originally designed for conformational sampling of small proteins and adapted to provide a fully unbiased descriptor selection (without any prior filtering, such as discarding correlated candidate descriptors) out of up to 10000 candidates. In order to avoid sampling of fictitious correlations that are likely to appear with increasing candidate descriptor sets, the fitness function already includes a leave-a-third-out cross-validation procedure. Several different control mechanisms of the number of selected variables were introduced, in order to ensure that potential models with few variables have been very thoroughly searched for before more descriptors are allowed to enter. Furthermore, the algorithm includes the ability not only to select (1) or discard (0) any of the descriptors, but to suggest possible predefined nonlinear transformations to be applied to the descriptor prior its entry in the multilinear regression model relating it to activity. The resulting models therefore actually represent one-layer neural nets having synapse weights set by the regression procedure and built on hand of descriptors that best suit that nonlinear context, rather than the ones preselected for their good behavior in linear models.

Benchmarking studies based on several HIV protease inhibitor training sets were run both with fuzzy pharmacophore triplets and ISIDA fragment descriptors, and using several splitting schemes into training and validation subsets. The ability of models issued from the different descriptor selection strategies to correctly predict affinites of the validation compounds was monitored and showed that the autoamted procedures succesfully cope with variable picking out of thousands of candidates all while avoiding the pitfall of fictitious correlations.