Meta-QSAR: learning how to learn QSARs

Ivan Olier[1], Crina Grosan[2], Noureddin Sadawi[2], Larisa Soldatova[2], and Ross D. King[1]

[1] Manchester Institute of Biotechnology
University of Manchester, United Kingdom

[2] Department of Computer Science
University of Brunel, United Kingdom

1 Introduction

Quantitative structure activity relationships (QSARs) are functions that predict bioactivity from compound structure. Although almost every form of statistical and machine learning method has been applied to learning QSARs, there is no single best way of learning QSARs. Therefore, currently the QSAR scientist has little to guide her/him on which QSAR approach to choose for a specific problem.

The aim of this work is to introduce Meta-QSAR, a meta-learning approach aimed to learning which QSAR method is most appropriate for a particular problem. For the preliminary results presented here, we used ChEMBL[1], a public available chemoinformatic database, to systematically run extensive comparative QSAR experiments. We further apply meta-learning in order to generalise these results.

2 Data and Methods

The datasets involved in this research have been formed by computing molecular properties and fingerprints of chemical compounds with associated bioactivity to a particular target (protein). Learning a QSAR model consists on fitting a regression method to a dataset which has as input variables the descriptors, as response variable (output) the associated bioactivities, and as instances, the chemical compounds. We extracted 2,750 targets from ChEMBL with a very diverse number of chemical compounds, ranging from 10 to about 6,000. Two sets of properties – one, using 43 constitutional properties, and another, using 1,683 additional properties – and one fingerprint (FCFP4, 1024bits) were used to form the datasets. Further datasets were generated by imputing missing values using the median and performing feature selection based on the chi-squared test. For the QSAR methods, we have selected 20 algorithms typically used in QSAR experiments, which include: linear regression, support vector machines, artificial neural networks, regression trees, and random forest, amongst others. Model performance in all experiments has been assessed by taking

[1] ChEMBL database is available from: https://www.ebi.ac.uk/chembl/

the average root mean squared error (RMSE) after 10-fold crossvalidation of the datasets.

For the meta-learning stage, we conceived a classification problem that indicates which QSAR method should be used for a particular QSAR problem. The training and learning dataset is formed by meta-features extracted from the datasets of the base learning level and are based on target properties (hydrophobicity, molecular weight, aliphatic index, etc) and on information theory (mean, mutual information, entropy, etc). We used random forests as meta-learning algorithm.

3 Results and Discussion

Fig. 1 shows preliminary results of the experiments. The graph on the left confirms the hypothesis that there is no single way to learning QSARs. Random forests proofs successful for the FCFP4 fingerprint representation, although other QSAR methods had good performance, too. The graph on the right is an evidence of the fact that Meta-QSAR learning is correctly suggesting for almost all targets which QSAR method should be used.

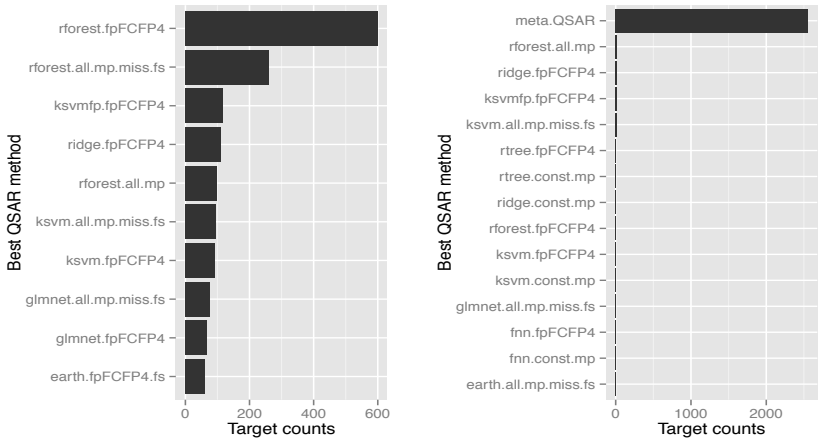

Fig. 1. Graphical representation of the number of times (target counts) a particular QSAR learning method obtains the best performance (minimum RMSE). Left: Results from the QSAR experiments. Right: Results using Meta-QSAR. The method names follow this convention: first term indicates the algorithm ('rforest', random forest; 'ksvm', support vector machine (SVM) with radial basis functions kernel; 'ksvmfp'; SVM with Tanimoto kernel; 'ridge', linear regression with ridge penalisation term; 'glmnet', elastic-net regularized generalised linear model; 'earth', multivariate adaptive regression splines; 'rtree', regression trees; 'fnn', fast k-Nearest Neighbour), second term, the kind of chemical compound descriptor set ('fpFCFP4', FCFP4 fingerprint; 'all.mp', full set of molecular properties; and 'const.mp', constitutional set of molecular properties), and then, optionally, whether missing value imputation ('miss') and feature selection ('fs') methods were used.