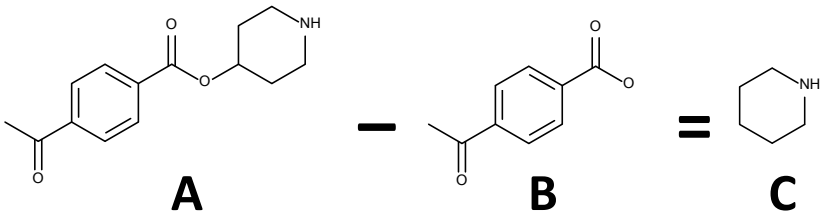
[P19] Knowledge mining from chemical datasets based on interpretation of QSAR models

Pavel Polishchuk

Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacký University and University Hospital in Olomouc, Hněvotínská 1333/5, 779 00 Olomouc, Czech Republic

Knowledge about structure-activity relationship is fundamental in chemistry and related fields. It is widely used for rational design of new compounds with desired properties. All approaches of SAR rules extraction can be divided on two groups. Approaches of the first group are based exclusively on compound structures and associated property values: frequent patterns, emergent and jumping emerging patterns, matched molecular pairs, etc. They are fast and simple but can be applied mainly to big datasets and classification problems and they are focused on explanation rather than prediction. The second group of approaches are based on extraction of SAR rules from computational models: QSAR, pharmacophore, molecular docking, etc. They are more complex but building of predictive models creates confidence that structure-activity relationship is observed and models captured it. These approaches can work with datasets of any size. These make interpretation of computational models an attractive alternative in knowledge mining from chemical datasets.

Recently the universal approach to structural interpretation of QSAR models was proposed which can be applied for any QSAR models regardless used machine learining method and descriptors [1]. We extended this approach and made it possible to estimate not only overall fragment contributions but contributions of physico-chemical factors as well. The approach can be illustrated by the following scheme :



Interpretation	$Activity_{pred}(A)$	$Activity_{pred}(B)$	Contribution(C)
Structural	$f(A_E, A_H, A_D, A_{HB}) = x$	$f(B_E, B_H, B_D, B_{HB}) = x$	$W(C) = x - y$
Physico-chemical	$f(A_E, A_H, A_D, A_{HB}) = x$	$f(A_E, A_H, A_D, B_{HB}) = x$	$W_{HB}(C) = x - y$

where A is a compound of interest, C – the fragment of interest whose contribution is to be calculated, B – the remaining part of A after removal of C (counter-fragment). f() is a QSAR model. W(C) – overall contribution of the fragment C. A_E, A_H, A_D, A_{HB} – descriptors of the compound A representing electrostatic, hydrophobic, dispersive and hydrogen bonding terms, correspondingly. B – descriptors of the compound B. $W_{HB}(C)$ – contribution of the fragment C regarding hydrogen bonding effects.

The approach was implemented as the open-source standalone software tool SPCI with GUI and available on http://qsar4u.com/pages/sirms_qsar.php. It was successfully applied towards different datasets and end-points and demonstrated its ability to capture known relationships as well as to find the new ones. It can also be used for finding of possible structural alerts and estimation of main physico-chemical factors to uncover the nature of underlying interactions.

The project was supported by the National program of sustainability LO1304.

Bibliography:

[1] Polishchuk, P.G.; Kuz'min, V.E.; Artemenko, A.G.; Muratov, E.N. Mol. Inf. 32 (2013) 843-853.