MODELING ROBUST QSAR

Jaroslaw Polanski

Department of Organic Chemistry, Institute of Chemistry, University of Silesia,
PL-40-006 Katowice, Poland, e-mail: polanski@us.edu.pl

It may look like a paradox but *the most fundamental and lasting objective of* (chemical) *synthesis is not a production of new compounds but the production of properties* [1]. Molecular design is a computational tool for screening virtual chemical compound space in a search for novel properties, and QSAR should work like a dictionary between molecular structures and properties. This clearly makes it an essential and irreplaceable method in molecular design. However, more and more sophisticated and robust tools are needed for the efficient transformation of the molecular structure space into the compound property space.

Molecular superimposition is a first problem during any QSAR procedure. Even if we use a method that does not require this operation, in fact, it is realized by default. This is convenient but we cannot control it. Although we usually do not realize this, in fact, such default superimposition is also performed by a variety of 2D QSARs. Recently, several improvements in the structure overlay have appeared that allow for more flexible or sophisticated superimposition.

Coding molecules in 3D or 4D QSAR is a next issue that does influence final model robustness. In the CoMFA-like fields a molecule is represented by a set of points determined in the space by a 3D grid. Different smooth and box fields have been thoroughly tested recently. A surface can be a base for the molecule description in several methods, e;g., Compass, CoRSA or CoMSA, that are based on sampling points or surface sectors. In Hopfinger's 4D QSAR a molecule is coded by the descriptors defining the pattern in which atoms occupy volume sectors. Alternatively, the self-organizing neural network can be used for the generation of molecular volumes.

Data handling is a next issue that can improve QSAR robustness. New computational methods including neural networks, data elimination, genetic algorithms, novel model validation schemes are some examples in this field. Generally, during QSAR modeling we operate on a strictly finite set of molecules for which activity is measured and described *a priori*. Eventually, before calculation we must have chosen the appropriate data for the compounds that are active. This decides that QSAR is more an *a posteriori* analysis of the SAR data structure than a strict method for the activity prediction in a sense of a novel compound design. Even a minute chemical structure modification can result in substantial activity changes. This similarity paradox decides that a virtual molecule, in reality, cannot only be a more or less substantial outlier in QSAR equation but can appear completely inactive. Thus, instead modeling by equations we can use more versatile techniques such as clustering or visualization. This allows us to avoid a paradox of producing an excellent model that completely fails when prediction is attempted. Eventually, it is better to have a vague idea on the trends than an illusion of a proper prediction. Finally, completely novel methods e.g., binary QSAR for HTS data appeared. This allows us to generate a model including both active and inactive compounds.

[1] Kolb, H. C., Finn, M. G., Sharpless, K. B., *Angew. Chem., Int. Ed. Engl.,* 40, 2004 (2001).