QSAR modeling, *Quo Vadis*?

Eugene Muratov, Denis Fourches, Alexander Tropsha

Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC 27599, USA

Keywords: *QSAR, cheminformatics, computer-aided molecular design, data curation, predictive modeling, QSAR of mixtures, QSAR of (nano)materials*.

Outline

Quantitative Structure-Activity Relationship (QSAR) modeling is one of the major computational tools to predict biological activity from the knowledge of chemical structure. Despite of its popularity, QSAR saw both praise and criticism concerning its reliability, limitations, successes, and failures.[1] In this presentation, we will discuss the current trends, unsolved problems, pressing challenges as well as several novel and emerging applications of QSAR modeling. We will describe our best practices for QSAR model development, validation, and application. These guidelines are primarily centered on the four following elements of our predictive QSAR workflow:

1.Data curation: Careful curation of chemical data is critical for the success of any cheminformatics analysis. We will present the workflows for both chemical and biological data curation[2] developed in our group at UNC. Treatment of chemical data includes the removal of inorganics, organometallics, counterions, and mixtures, structural cleaning (*e.g.,* detection of valence violations), ring aromatization, normalization of specific chemotypes, standardization of tautomeric forms, deletion of duplicates, and manual checking of complex cases. Furthermore, biological data curation encompasses the detection and verification of activity cliffs, analysis of experimental variability, calculation and tuning of dataset modelability index, and identification and correction of misannotated compounds based on predictions by QSAR.

2. MODelability Index (MODI): we will describe the concept of dataset modelability by QSAR that we introduced recently[3]. It was proposed not only as a quantitative tool to quickly estimate whether predictive QSAR model(s) could be obtained for a given dataset but also as an attempt to answer the following questions: *(i)* how the number of activity cliffs correlates with the overall prediction performance of QSAR models for a given dataset; *(ii)* is such correlation conserved across different datasets; *(iii)* can one use the fraction of activity cliffs in a datasets to assess the overall possibility of success or failure for QSAR modeling; *(iv)* why some datasets are modelable whereas others are not; and *(v)* is it possible to find within a non-modelable dataset, a subset of compounds for which local QSAR models could be obtained.

3.Model building and validation: We will describe the predictive QSAR modeling workflow developed in our lab with a particular attention to both internal and external cross-validation, estimation of applicability domain of QSAR models, and the generation of consensus predictions[4]. A brief overview of most popular molecular descriptors and machine learning-techniques will be given.

4.Applications: Experimental validation is the ultimate indicator of the predictive abilities of any QSAR model. We will show several examples of experimentally-assisted computational drug design including the development of novel compounds with the desired polypharmacological profiles as well as more traditional application of QSAR to the optimization of novel antivirals and antimicrobials. We will also describe non-trivial applications and future trends of QSAR such as modeling of peptides and chemical mixtures, quantitative nanostructure-activity relationships (QNAR), and the use of QSAR models in materials informatics.

Acknowledgements

The authors are thankful for IBM and UNC (Junior Faculty Development Award) for financial support of Dr. Muratov's research. The authors also appreciate the support provided by NIH (grants GM096967 and GM066940).

---

[1] Cherkasov, A.; Muratov, E.; Fourches, D; et al. *J. Med. Chem.* **2014**, *DOI: 10.1021/jm4004285*.
[2] Fourches, D.; Muratov, E.; Tropsha, A. *J. Chem. Inf. Model.* **2010**, *50*, 1189.
[3] Golbraikh, A.; Muratov, E.; Fourches, D.; Tropsha, A. *J. Chem. Inf. Model.* **2014**, *DOI: 10.1021/ci400572x*.
[4] Tropsha, A. *Mol. Info.* **2010**, *29*, 476.