Chemometrics and Intelligent Laboratory Systems 145 (2015) 22–29

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

On a simple approach for determining applicability domain of QSAR models

Kunal Roy [a,b,*], Supratik Kar [a], Pravin Ambure [a]

[a] Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India
[b] Manchester Institute of Biotechnology, University of Manchester, Manchester M1 7DN, United Kingdom

ARTICLE INFO

Article history:
Received 27 February 2015
Received in revised form 11 April 2015
Accepted 16 April 2015
Available online 25 April 2015

Keywords:
Applicability domain
Leverage
OECD
Outlier
QSAR
Standardization

ABSTRACT

Quantitative structure–activity/property/toxicity relationship (QSAR/QSPR/QSTR) modeling has been used in medicinal chemistry, material sciences, environmental fate modeling, risk assessment and computational toxicology for a long time. The Organization for Economic Co-operation and Development (OECD) has recommended that for application of validated QSAR models for prediction of new data points, there is a strict requirement of defining the applicability domain (AD) according to the *Principle 3*. The AD is a theoretical region in chemical space encompassing both the model descriptors and modeled response which allows one to estimate the uncertainty in the prediction of a particular compound based on how similar it is to the training compounds employed in the model development. The AD is an important tool for reliable application of QSAR models, while characterization of interpolation space is significant in defining the AD. An attempt is made here to suggest a simple method for defining the X-outliers (in the case of the training set) and identifying the compounds that reside outside the AD (in the case of the test set) employing the basic theory of the standardization approach. Further, a standalone application named "Applicability domain using standardization approach" (available at http://dtclab.webs.com/software-tools and http://teqip.jdvu.ac.in/QSAR_Tools/) has been developed. The present study reports that the web application can be easily used for identification of the X-outliers for training set compounds and detection of the test compounds residing outside the AD using the descriptor pool of the training and test sets.

1. Introduction

Quantitative structure–activity/property/toxicity relationship (QSAR/QSPR/QSTR) modeling has become an extensively used tool in computer-aided drug design (CADD), predictive environmental risk assessment and fate modeling, toxicity and property prediction of chemicals and pharmaceuticals [1,2] as well as in different modeling problems in material sciences [3], analytical chemistry and pharmacokinetics/pharmacodynamics profiling of new drug molecules [4]. A QSAR model is a simple mathematical relation derived from a set of training molecules with known activities/properties/toxicities using regression or classification based approaches. This technique offers an in silico tool for the development of predictive models towards various activity and property endpoints of a series of chemicals using the response data and molecular structure information derived computationally or obtained from experiments. QSAR is an economical and time-effective alternative to the medium throughput in vitro and low throughput in vivo assays [5]. The QSAR modeling also supports the 3R (replacement, refinement and reduction in animals in research) paradigm due to an increased socio-economic pressure to trim down the use of animal testing as an important alternative method for future prediction of untested chemical entities [6]. Most importantly, regulatory agencies worldwide have been employing a large number of QSAR models for decision-making frameworks in risk and safety assessments over the years [2].

A large numbers of studies have been directed to the design of new drugs with the utilization of the QSAR technique, and validation plays an important role in the development of predictive QSAR models which may be considered for the future prediction of new molecules [7]. As a consequence, investigations are currently directed towards the introduction of more apposite validation approaches for accurate and predictive QSAR model development. In this perspective, one important objective of QSAR modeling is to predict activity/property/toxicity of new chemical entities falling within the applicability domain (AD) of the developed model. The reliability of any QSAR model depends on the confident predictions of these new compounds based on the AD of the model, and therein lies the importance of the AD study [8]. However, even after approximately 50 years of QSAR research, many researchers are still unaware of the importance of defining the

* Corresponding author at: Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India. Tel.: +91 98315 94140; fax: +91 33 2837 1078.
E-mail addresses: kunalroy_in@yahoo.com, kroy@pharma.jdvu.ac.in, kunal.roy@manchester.ac.uk (K. Roy).
URL: http://sites.google.com/site/kunalroyindia/ (K. Roy).

http://dx.doi.org/10.1016/j.chemolab.2015.04.013
0169-7439/

applicability domain of the developed models for their useful application on new data points.

In order to strengthen the scientific validity of a QSAR model and to assist its acceptance for regulatory purposes, the Organization for Economic Cooperation and Development (OECD) in its joint meeting [9] has agreed to five principles that should be followed during the construction of QSAR models. The OECD Principle 3 defines '*a defined domain of applicability*' for the developed QSAR model. The Setubal Workshop report [10] presented the following regulation for the AD assessment: "*The applicability domain of a (Q)SAR is the physico-chemical, structural, or biological space, knowledge or information on which the training set of the model has been developed, and for which it is applicable to make predictions for new compounds. The applicability domain of a (Q)SAR should be described in terms of the most relevant parameters, i.e., usually those that are descriptors of the model. Ideally the (Q)SAR should only be used to make predictions within that domain by interpolation not extrapolation.*" This illustration is helpful for explaining the intuitive meaning of the "applicability domain" approach. The AD of a QSAR model has been defined as the response and chemical structure space, characterized by the properties of the molecules in the training set. The developed QSAR model can predict a new compound truly only when it falls within the AD of the constructed model [11]. Thus, identifying the interpolation (true prediction) or extrapolation (less reliable prediction) of query compounds is a vital task for a QSAR model developer using the information of the AD [11].

The domain of applicability of molecules plays a critical role for estimating the uncertainty in the prediction of a specific molecule based on how similar it is to the compounds employed to construct the model. Thus, the prediction of a modeled response using QSAR is valid only if the compound being predicted falls within the AD of the model as it is impractical to predict a whole universe of chemicals employing a single QSAR model. Again, the selection method of the training and test sets has a significant impact on the QSAR model as there is a high possibility of considering outliers in the training set (which are actually influential observations for the model) and/or including compounds much dissimilar to the training set compounds in the test set. Thus, while splitting a dataset for external validation, the training set molecules should be selected in such a way that they span the entire chemical space for all the dataset molecules with proper handling of outliers. On the contrary, if a new compound falls outside of the AD of the training set molecules, its prediction is not reliable.

Viewing the importance of AD in the application of QSAR models, we have attempted to propose a simple method for defining outliers (in the case of the training set) and the compounds residing outside the AD (in the case of the test set) to build reliable and acceptable QSAR models employing the basic theory of the *standardization* approach. Further, a standalone application named "Applicability domain using standardization approach" (available at http://dtclab.webs.com/software-tools and http://teqip.jdvu.ac.in/QSAR_Tools/) has been developed. We may mention here that development of online tools to newly proposed methodologies is of increasing practice in recent years as it has various advantages including worldwide access to the new methodologies [12]. Apart from this, freely accessible and highly efficient workflow systems are also being introduced by different research groups in order to incorporate various QSAR functionalities including calculation of applicability domain. For instance, KNIME (Konstanz Information Miner) workflow system can be used to compute applicability domain using two Enalos KNIME Nodes, i.e., Enalos Domain –*Similarity node that can be used to define Applicability Domain (APD) based on the Euclidean distances* and Enalos Domain – *Leverages node that can be used to define Applicability Domain based on the Leverages* [13,14].

2. Types of available methods for determining AD

There are various approaches for determining AD of QSAR models. The most commonly employed approaches for estimating the interpolation regions in a multivariate space include the followings [10,11,15]:

1. Ranges in the descriptor space.
2. Geometrical methods.
3. Distance-based methods.
4. Probability density distribution.
5. Range of the response variable.

Here, the first four approaches are based on the methodology used for interpolation space characterization in the model descriptor space. On the contrary, the last one depends solely on response space of the training set molecules. The existing methods for determining AD are depicted in Fig. 1. Among the existing methods, no method can be considered as the universally best. Each method has its own merits and flaws. From the QSAR publications of the last decade, one can see that the leverage approach (Williams plot) [16], DModX [17] and different similarity assessment approaches [15] have been largely employed to identify the outliers and compounds residing outside the AD. These methods are dependent on the usage of specialized statistical software tools to some extent for the calculation of different parameters for outlier identification and determination of AD. In the presented work, we have attempted to report a simple statistical approach to define AD of a QSAR model and also to develop a software tool implementing the method which is freely downloadable from the Web.

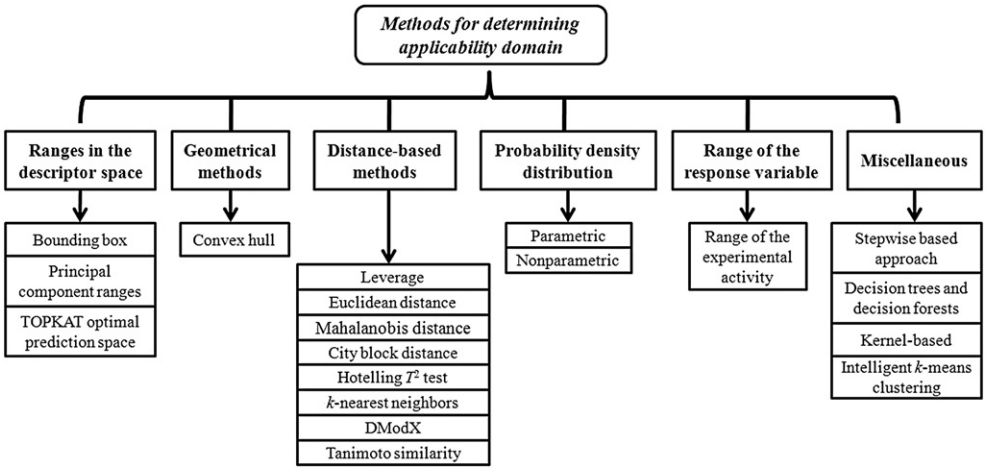

Fig. 1. The existing methods for determining applicability domain.