J. Indian Chem. Soc.,
Vol. 95, December 2018, pp. 1497-1502

Quantitative structure-activity relationships (QSARs): A few validation methods and software tools developed at the DTC laboratory†

Kunal Roy

Drug Theoretics and Cheminformatics (DTC) Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata-700 032, India

E-mail: kunal.roy@jadavpuruniversity.in *Fax:* 91-33-28371078

Manuscript received online 02 November 2018, accepted 26 November 2018

In this presentation, different quantitative structure-activity relationship (QSAR) modeling approaches and their use in drug design and ecotoxicological modeling are briefly stated. The aspects of feature selection, modeling algorithms and validation strategies are mentioned at an elementary level. Different novel strategies for improving statistical quality and predictive ability of QSAR models are also cursorily presented. Finally, four useful tools for QSAR model validation as developed by the Drug Theoretics and Cheminformatics (DTC) Laboratory of Jadavpur University are discussed. These tools are available for public use via http://teqip.jdvu.ac.in/QSAR_Tools/ and https://dtclab.webs.com/software-tools.

Keywords: QSAR, validation, software tools, DTC laboratory.

Introduction

Quantitative structure-activity relationships (QSARs) are statistical models, which can be developed based on a similarity principle to correlate the changes in the biological activity (or property or toxicity) of chemicals (including pharmaceuticals, cosmetics, agrochemicals, nanomaterials, and so on) with changes in their structural features[1–5]. Such changes in the biological activity or other property of chemical compounds occur in a systematic way with the changes in the structural features or other physicochemical properties making it possible to develop quantitative mathematical models for structure-activity correlations. QSARs have long been applied in drug design and predictive toxicology in addition to their more recent applications in materials science, food sciences, nanosciences, etc.[5,6]. These models are used mainly for two purposes, prediction of the endpoint values for untested chemicals for data gap filling, and physico-chemical and mechanistic interpretations of the structure-response relationships. In general, classical QSAR approaches are more efficient for the second purpose while more recent machine learning and intelligent methods are more useful for the first purpose[7]. QSARs may be regarded as a subdiscipline of the broader area of Cheminformatics, and in association with other ligand-based (such as pharmacophore) and structure-based (such as molecular docking) approaches, these models may be very helpful in the design of novel compounds with optimum activity profile and screening of virtual libraries[8]. In order to have precise quantitative predictions, a regression-based approach may be used while for class-wise or graded qualitative predictions, a classification-based approach might be used. To apply statistical methods, it is necessary to have the structural (and property) information available in the form of numbers, which are termed descriptors. While descriptors can be readily computed or derived from chemical structure or property upto the 2D level, 3D descriptors require additional complexity in computation in terms of conformational analysis and energy minimization of chemical structures. Due to availability of a plethora of descriptors[9], which can be readily be computed using various available software tools, it is also important to apply a descriptor-thinning process followed by a feature selection tool before applying the modeling method[10]. For feature selection, methods like stepwise selection, genetic method, factor analysis etc. have been used in the QSAR literature[10]. Among the regression-based modeling techniques, multiple linear regression, partial least squares, principal component regression

†**Professor R. D. Desai 80th Birthday Commemoration Lecture (2017).**

analysis, ridge regression etc. have been used. Among the classification-based techniques, one can use linear discriminant analysis, *k*-means cluster analysis, etc. Various machine learning tools like deep neural network, support vector machine, random forests, etc. are very popular in the current QSAR research[1].

QSAR models are extensively used for regulatory purposes in chemical industries of the European Union (EU) in view of the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) regulations and other EU regulations specific to particular uses of chemicals like cosmetics, pigments, biocides, etc. QSARs can be used to predict the environmental properties of chemicals against several endpoints for which experimental data are not available[11–13]. The regulatory QSAR models should be developed based on five point guidelines laid down by Organization for Economic Co-operation and Development (OECD) for QSAR model development and validation[14]. These guidelines recommend a defined endpoint for modeling (ensuring same experimental protocol for the endpoint values), an unambiguous algorithm for model development (which ascertains reproducibility), a defined chemical applicability domain of the model (which ensures that the query chemicals are sufficiently similar to the compounds used for model development), appropriate use of statistical measures for checking fitness and predictive ability of the developed model (which decides the acceptability of a model) and finally, mechanistic interpretability of the model, if possible. It is necessary to apply a variety of statistical methods and metrics (depending on the regression-based or classification-based modeling methods being used) to examine the statistical quality of the developed models. However, the model quality itself does not ascertain the prediction quality for which one has to apply validation tools encompassing internal validation (including leave-one-out and leave-many-out cross-validation), Y-randomization and external validation tools[15]. The predictive quality of models is expressed in terms of various validation metrics and their recognized threshold values[16,17]. For regression-based QSAR models, the quality metrics may in general be classified into two groups, correlation based metrics and error based metrics while the latter type gives a more direct information about model quality[18]. It is also important to analyze for any bias in prediction errors in order to identify any systemic error in the model[19]. In case of classification-based models, the quality metrics are derived from a contingency matrix. In general, external validation has been considered as the gold standard for checking predictive ability of models for new chemicals[15]. For this purpose, the available data set is usually divided into two parts, a training set, which is used for model development, and a test set, which is kept aside and never used for descriptor selection and model development. The test set is subsequently used for checking the quality of predictions from the developed model. Thus, the performance of a QSAR model is always checked against the experimental values of new or test compounds. The division of the whole data set into training and test sets should be done based on a principle of similarity[20] while taking into account the optimum size of the training set required for the learning process[21]. It is also very important to check the applicability of a QSAR model for a new chemical, which may actually be outside the domain of the model due to its structural difference to the training set molecules[22,23]. The concept of overfitting is also very important while development of QSAR models. Thus, one should restrict using a higher number of descriptors in the model to avoid limited degrees of freedom leading to an overfitted model. However, the issue of the number of descriptors may be less relevant for more robust and machine learning tools, which can handle very complex data, but the removal of noisy descriptors always increase the predictive ability of models. Among the emerging trends in QSAR analysis, we may mention here multi-target and multi-task QSAR analyses[29,30], which would be very promising in coming days.

There are several prerequisites for the preparation of experimental data ready for QSAR analysis[1]. Usually the concentrations or doses required for a fixed response such as EC_{50}, ED_{50}, IC_{50} or LD_{50}values are used as the response for activity or toxicity based QSAR analyses. The concentration values are expressed in a molar unit and expressed in a negative logarithmic scale so that a higher value represents higher activity or toxicity. The logarithmic conversion can also handle a wide span of concentration values apart from the linearity of the log dose-response curve over an extended range. It should be remembered that all QSAR models are derived from statistical treatments, and they are not mathematical solutions. Thus, there should be a good degree of freedom to ascertain statistical soundness of a QSAR model. Therefore, the number of observations based on which a

model is developed should be considerably high with respect to the number of descriptors (constraints) used in the modeling. Although this aspect is less important for more robust techniques, the use of sufficient number of training compounds cannot be ignored even in case of machine learning techniques. The issue of modeling with small data sets is really a big problem in QSAR research, as for several endpoints, sufficient number of experimental observations might be unavailable. Multiple linear regression (MLR) is a commonly used method for activity and toxicity based classical type QSARs while it presents several problems like intercorrelation among descriptors, bias in descriptor selection due to a fixed composition of the training set, inability to handle many descriptors in the model, etc. This problem may be overcome by using a more robust modeling technique like partial least squares (PLS)[24], which converts the original set of descriptors into a lower number of latent variables which are functions of the original descriptors. However, the dataset with a small number of data points needs a special attention during modeling. A double cross-validation technique[25,26] may be of help in such cases. In this approach, the validation is done in two loops: in the inner loop, the training set is further divided into 'n' calibration and validation sets resulting in diverse compositions, which are utilized for model building and model selection, while the test set in the external loop is exclusively used for model assessment. In another approach, consensus predictions have been applied in several studies as more reliable than individual model derived predictions, as the former takes into account contribution of maximum possible combination of important descriptors. The final result considers the different assumptions characterizing each method for a more reliable judgment in a complex situation. This approach can also afford greater chemical space coverage. Recently, an intelligent consensus modeling method has been reported considering that a single QSAR model may not be equally good for predictions for all query compounds[27]. It is also important to evaluate the reliability of predictions for untested compounds, which may not be dependent solely on applicability domain. There are different approaches in the literature, but we mention here the tool "Prediction Reliability Indicator" tool for MLR and PLS predictions[28].

In the Drug Theoretics and Cheminformatics (DTC) Laboratory of Jadavpur University, we have developed several methods for QSAR model validation[18,23,27,28]. We have also developed software tools for MLR and PLS model development and validation such as MLR plus validation, XternalValidationPlus, and Partial Least Squares. Apart from these tools, there are several small tools, which are also very useful before and during QSAR model building such as Normalize Data, Data Pretreatment, Stepwise MLR, Genetic Algorithm, etc. All these software tools have been developed in Java by Pravin Ambure (ambure.pharmait@gmail.com; Present Affiliation: FACULDADE DE CIÊNCIAS DA UNIVERSIDADE DO PORTO, Portugal) and they have been made available for public use free of cost via http://teqip.jdvu.ac.in/QSAR_Tools/ and https://dtclab.webs.com/software-tools. Here, we present four recent additions to this collection of tools.

1. Applicability domain using standardization approach

Applicability Domain (AD) is defined as "the response and chemical structure space in which the QSAR model makes predictions with a given reliability". The "AD using Standardization approach" is a tool[23] to detect outliers from training set compounds and find out test compounds that are outside the applicability domain. The basic principle applied in this approach is as follows:

A QSAR model is trained from the features present in the training set compounds. The developed model is then applied for prediction of test set compounds, which should ideally be structurally similar to the training set compounds,as the model has captured similar features present in the training set. If a small fraction of the training is very dissimilar to the rest and most of the compounds, then obviously those features are not properly included in the training process. These compounds are *X*-outliers. If test set compounds are similar to these small fraction of training set compounds, then their predictions are expected to be unreliable, as the model has not captured the features of those training set compounds, which have a small representation and are different from majority of the compounds. Therefore, these test set compounds are expected to be outside the AD of the model. Again, the test set compounds, which are not similar to any of the training set compounds are also outside the AD.

Ideally, all the descriptors of the training set compounds should follow a normal distribution pattern. According to this distribution, 99.7% of the population will remain within the range mean ±3 standard deviation (SD). Thus, mean ±3SD represents the zone which most of the training set compounds

correspond to. Any compound outside this zone is dissimilar to the rest and majority of the compounds. Thus, after a descriptor column is standardized based on the corresponding mean and standard deviation for the training set compounds only, if the corresponding standardized value for descriptor *i* of compound *k* (S_{ki}) is more than 3, then the compound should be an *X*-outlier (if it is in the training set) or outside AD (if it is in the test set) based on descriptor *i*. This test should run for all descriptors present in the model. If the maximum S_i value of a compound *k* is lower than 3, then the compound is quite similar to a good number of compounds in the training set with respect to all descriptors (not an *X*-outlier if in the training set and is within AD if in the test set). If the minimum S_i value of a compound *k* is higher than 3, then the compound is quite dissimilar to most of the compounds in the training set with respect to all descriptors (an *X*-outlier if in the training set and not within AD if in the test set). If the compound has a maximum S_i value above 3 but the minimum S_i value is below 3, then the compound is similar to most of the training set compounds with respect to some descriptors and at the same time dissimilar to most of the training set compounds with respect to other descriptors. Thus, we need an additional criterion of assessment of *X*-outliers or applicability domain behavior of such compounds. Now again considering an ideal case of standardized normal distribution, the standard score (*Z*) corresponding to 1.28 represents a relative frequency of occurrence of less than 1.28 times SD being 90%. Thus, in our case, if mean of the S_i values of a compound for all descriptors in a model plus 1.28 times corresponding standard deviation (termed as S_{new}) is lower than 3, then there is 90% probability that the S_i values of that compound are lower than 3. Thus, when S_{new} value of a compound is lower than 3, then the compound can be considered to be not an X-outlier (if in the training set) or within the AD (if in the test set). This assumption is statistically more valid when a higher number of descriptors are present in the model. The dedicated web page for this tool is https://sites.google.com/site/dtclabad/.

2. Double cross-validation

The double cross-validation process comprises two nested cross-validation loops.These are referred to as internal and external cross-validation loops. In the outer (external) loop of double cross-validation, all data points are divided into two subsets referred to as training and test sets. The training set is used in the inner (internal) loop of double cross-validation for model building and model selection, while the test set is exclusively used for model assessment. In the internal loop, the training set is repeatedly split *k* times into calibration and validation data sets. The calibration objects are used to develop different models whereas the validation objects are used to estimate the models' error. Finally, the model with the lowest prediction errors (validation set) in the inner loop is selected. Then, the test objects in the outer loop are employed to assess the predictive performance of the selected model. This method of multiple splits of the training set into calibration and validation sets obviates the bias introduced in variable selection in case of usage of a single training set of fixed composition.

The "Double crossvalidation" tool[26] performs MLR model development using the double cross-validation technique as mentioned above. Optionally, this tool can also simultaneously develop PLS models. Further, it also provides two variable selection techniques (stepwise method and genetic algorithm) and four different ways of selecting the optimum model (consensus predictions, and methods based on the least mean absolute error of the validation set, based on best subset selection in case of MLR and using pooled descriptors for PLS). The dedicated webpage for this tool is https://sites.google.com/site/dtclabdcv/.

3. Intelligent consensus predictor

The "Intelligent Consensus Predictor" tool[27] judges the performance of "intelligent" consensus predictions obtained from multiple QSAR models (MLR or PLS) developed against a particular response and compares them with the prediction quality obtained from the individual models. This tool performs four different ways of consensus predictions along with the individual model predictions. The basic assumption for this tool is that a single model may not be equally good for predictions for different query compounds. Further, the quality of predictions is judged based on several external validation metrics. Moreover, this tool also provides few optional criteria (i.e., Euclidean distance cut-off, applicability domain and Dixon-Q test) that might help in improving the quality of prediction for a query molecule by considering the aspects such as the level of chemical similarity, prediction outliers in training compounds, etc. The optimum settings

can be fixed using the available QSAR models and corresponding external set compounds with known response values, while the same setting can later be employed for predictions of newly designed query molecules. The dedicated webpage for this tool is https://sites.google.com/site/dtclabicp/.

4. Prediction reliability indicator

The tool "Prediction Reliability Indicator"[28] was developed to indicate or categorize the quality of predictions for the test set or external set (with or without experimental or observed response (Y) values) into three groups: good (*with composite score 3*), moderate (*with composite score 2*) and bad (*with composite score 1*). We have used here three different criteria in different weighting schemes for making a composite score of predictions: *(1) mean absolute error of leave-one-out predictions for 10 most close training compounds for each query molecule; (2) applicability domain in terms of similarity based on the standardization approach; (3) proximity of the predicted value of the query compound to the mean training response*. The tool can automatically find the optimum weightage based on %correct predictions computed using a test set with known observed response and thus known quality of predictions. However, the user also has an option to select the weightage manually, especially when the experimental response values (Y) are unknown. The dedicated webpage for this tool is https://sites.google.com/site/dtclabpri/.

The above tools for QSAR model validation should be useful for the QSAR community as evidenced from the combined citation of the related four recent papers[23, 26–28] being 203 on SCOPUS (https://www.scopus.com/) as on December 03, 2018.

Acknowledgement

Financial assistance received from different funding agencies like University Grants Commission, New Delhi, All India Council for Technical Education, New Delhi, Department of Biotechnology, New Delhi and Indian Council of Medical Research, New Delhi over different time periods is thankfully acknowledged.

References

1. K. Roy, S. Kar and R. N. Das, "Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment", Academic Press, New York, 2015.
2. K. Roy, S. Kar and R. N. Das, "A Primer on QSAR/QSPR Modeling: Fundamental Concepts (Springer Briefs in Molecular Science)", Springer, New York, 2015.
3. K. Roy (Ed.), "Quantitative Structure-Activity Relationships in Drug Design, Predictive Toxicology, and Risk Assessment", IGI Global, PA, 2015.
4. K. Roy (Ed.), "Advances in QSAR Modeling. Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences", Springer, New York, 2017.
5. J. C. Dearden, *Int. J. Quant. Struct.-Prop. Relat.*, 2016, **1(1)**, 1.
6. A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard and A. Tropsha, *J. Med. Chem.*, 2014, **57(12)**, 4977.
7. T. Fujita and D. A. Winkler, *J. Chem. Inf. Model*, 2016, **56(2)**, 269.
8. G. I. Passeri, D. Trisciuzzi, D. Alberga, L. Siragusa, F. Leonetti, G. F. Mangiatordi and O. Nicolotti, *Int. J. Quant. Struct.-Prop. Relat.,* 2018, **3(1)**, 134.
9. R. Todeschini and V. Consonni, "Handbook of Molecular Descriptors", Wiley-VCH, Weinheim, 2008.
10. P. M. Khan and K. Roy, *Expert Opin. Drug Discov*., 2018, http://dx.doi.org/10.1080/17460441.2018.1542428.
11. M. T. D. Cronin, *Environ. Sci.: Processes Impacts*, 2017, **19**, 213.
12. S. Kar and K. Roy, *J. Indian Chem. Soc.*, 2010, **87**, 1455.
13. R. Gozalbes and J. V.de Julián-Ortiz, *Int. J. Quant. Struct.-Prop. Relat.*, 2018, **3(1)**, 1.
14. http://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm (accessed November 09, 2018).
15. A. Tropsha, *Mol. Inf.*, 2010, **29**, 476.
16. K. Roy and I. Mitra, *Comb. Chem. High Throughput Screen.*, 2011, **14**, 450.
17. K. Roy, I. Mitra, S. Kar, P. K. Ojha, R. N. Das and H. Kabir, *J. Chem. Inf. Model*, 2012, **52**, 396.
18. K. Roy, R. N. Das, P. Ambure and R. B. Aher, *Chemom. Intell. Lab. Syst.*, 2016, **152**, 18.
19. K. Roy, P. Ambure and R. Aher, *Chemom. Intell. Lab. Syst.*, 2017, **162**, 44.
20. J. T. Leonard and K. Roy, *QSAR Comb. Sci.*, 2006, **25**, 235.
21. P. P. Roy, J. T. Leonard and K. Roy, *Chemom. Intell. Lab. Syst.*, 2008, **90**, 31.
22. D. Gadaleta, G. F. Mangiatordi, M. Catto, A. Carotti and O. Nicolotti, *Int. J. Quant. Struct.-Prop. Relat.*, 2016, **1(1)**, 45.
23. K. Roy, S. Kar and P. Ambure, *Chemom. Intell. Lab. Syst.*, 2015, **145**, 22.
24. S. Wold, M. Sjöström and L. Eriksson, *Chemom. Intell.*

Lab. Syst., 2001, **58(2)**, 109.

25. D. Baumann and K. Baumann, *J. Cheminformatics*, 2014, **6(1)**, 47.
26. K. Roy and P. Ambure, *Chemom. Intell. Lab. Syst.*, 2016, **159**, 108.
27. K. Roy, P. Ambure, S. Kar and P. K. Ojha, *J. Chemom.,* 2018, **32**, e2992.
28. K. Roy, P. Ambure and S. Kar, *ACS Omega*, 2018, **3**, 11392.
29. P. Ambure, J. Bhat, T. Puzyn and K. Roy, *J. Biomol. Str. Dyn.*, 2018, https://doi.org/10.1080/07391102.2018.1456975.
30. K. Roy (Ed.), "Multi-Target Drug Design Using Chem-Bioinfor-matic Approaches", Springer, 2019, https://www.springer.com/us/book/9781493987320.