European Journal of Pharmaceutical Sciences 62 (2014) 111–114

Contents lists available at ScienceDirect

European Journal of Pharmaceutical Sciences

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

Commentary

The r_m^2 metrics and regression through origin approach: Reliable and useful validation tools for predictive QSAR models (Commentary on 'Is regression through origin useful in external validation of QSAR models?')

Kunal Roy [a,b,*], Supratik Kar [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 6 May 2014
Received in revised form 21 May 2014
Accepted 21 May 2014
Available online 29 May 2014

Keywords:
QSAR
External validation
r_m^2 Metrics
Regression through origin

ABSTRACT

Quantitative structure–activity relationship (QSAR) is an *in silico* technique which can be used in drug discovery, environmental fate modeling, property and toxicity prediction of chemical entities and regulatory toxicology. The predictive potential of a QSAR model is judged from various validation metrics in order to evaluate how well it is capable to predict endpoint values of new untested compounds. The r_m^2 group of metrics is one of the stringent validation metrics currently used by the QSAR fraternity in different reports. We scrutinized a recently published paper which raised an issue that the constructed criteria based on regression through origin (RTO) are not optimal and there is a significant difference in the r_m^2 metrics values computed from different statistical software packages. According to our point of view, the conclusion drawn in this paper appears to be misleading. Any inconsistency in the software algorithms has nothing to do with the calculation of r_m^2 metrics, as such computation is not limited by the use of any specific software, rather it depends only on fundamental mathematical formulae that are well established. However, it is a concern to the QSAR users that Excel and SPSS can return different results for the metrics using the RTO method. Thus, a proper validation of the software tool is required before its use for computation of any validation metric.

1. Introduction

A great deal of recent research has been directed towards the modeling and design of new chemicals and pharmaceuticals worldwide (Helguera et al., 2008). The quantitative structure–activity relationship (QSAR) methodology is one of the most common and largely used computational tools which deals with the correlation between biological activity/toxicity/property of molecules and their structural features (Perkins et al., 2003). The QSAR models are particularly suitable for drug design, molecular modeling, chemical engineering problems and especially for decision-making frameworks in chemical safety assessment worldwide. QSAR also plays an important role in lead structure optimization in association with combinatorial chemistry (Kar and Roy, 2010).

Validation strategies are recognized as one the most decisive steps for the acceptability of any QSAR models for their future use on a new set of data for the confident predictions (Tropsha et al., 2003). There has been a great deal of debate regarding the preference of the most appropriate validation metrics for QSAR modeling among the QSAR researchers (Roy, 2007). The traditional internal and external validation metrics exhibit satisfactory results as long as good correlation is maintained between the observed and the predicted response data irrespective of the actual difference between the data. An alternative measure of r_m^2 (modified r^2) was suggested by Roy and coworkers (Mitra et al., 2010; Ojha et al., 2011; Roy et al., 2009, 2012; Roy and Roy, 2008) to be a better and more stringent metric for selection of the best predictive QSAR models.

The r_m^2 metrics depend chiefly on the difference between the observed and predicted response data and convey more precise information regarding their difference. Therein lies the utility of

* 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.ejps.2014.05.019
0928-0987/© 2014 Elsevier B.V. All rights reserved.

the $r_m{}^2$ metrics. The $r_m{}^2$ metrics analyze the models solely based on their ability to predict the activity of the training/test/overall set of compounds and thus facilitate an improved screening of the most predictive *in silico* models. QSAR researchers have revealed the significance of the $r_m{}^2$ metrics for the selection of the best QSAR models in their research work (Consonni et al., 2009; Gálvez-Llompart et al., 2011; Hu et al., 2009; Kar and Roy, 2013; Prado-Prado et al., 2009; Roy and Popelier, 2009; Toropov et al., 2010).

2. Background

We felt the need of writing a commentary due to some misleading arguments and hypotheses published in a recent paper by Shayanfar and Shayanfar (2014) where the authors raised a question "Is regression through origin useful in external validation of QSAR models?". The authors have mentioned in their paper that though the most widely used criteria for external validation which have been applied in hundreds of recent QSAR studies are the Golbraikh–Tropsha and Roy methods which are based on the regression through origin (RTO), but "*there is an inconsistency in the definition and calculation of r^2 of RTO*" (*i.e.*, r_0^2) and therefore "*the constructed criteria based on RTO is not optimal*".

They have tried to establish the above statement from the following sentences. "*There is a significant difference between the RTO formulae in Excel and SPSS statistical packages. Excel can give negative values of r^2 whenever intercept value is large and was estimated without intercept.*" They have calculated all of the statistical parameters (i.e. r^2, r_0^2, $r_0'^2$, k, k' and $r_m{}^2$) for 6 different models based on different datasets separately using Excel 2003 and SPSS 11.5 and the results of these statistical package for external validation of different data sets were compared. They have also mentioned in their results that "*A significant difference between calculated r_0^2 values by SPSS and Excel was observed. According to the Excel results, the difference between r^2 and r_0^2 and numerical values of K and K′ are acceptable, therefore the developed models are valid. On the other hand, the SPSS gave contradictory results. The value of r_0^2 is near 1 and it is not possible to calculate r_m^2 due to $r_0^2 > r^2$*".

The conclusion drawn in this paper (Shayanfar and Shayanfar, 2014) appears to be misleading and the objective of writing this paper is unclear as the method of calculation of r^2 of RTO (*i.e.*, r_0^2) has been described long ago in the text of applied statistics (Sachs, 1982) and its use is not limited by the use of any particular software. We may ask the authors: "Are you addressing the problem of a particular software or discussing the application of a validation metric?" One has to remember that the computation of $r_m{}^2$ metrics is not dependent on any specific software, it is dependent only on basic statistics.

3. Method of analysis

Roy and co-workers proposed the novel $r_m{}^2$ metrics as an important group of validation parameters (Roy et al., 2012). The $r_m{}^2$ metric is calculated based on the correlations between observed and predicted values with (r^2) and without (r_0^2) intercept for the least squares regression lines as shown in the following equation:

$$r_m^2 = r^2 \times \left(1 - \sqrt{(r^2 - r_0^2)}\right) \tag{1}$$

The squared correlation coefficient values between the observed and predicted values with intercept (r^2) and without intercept (r_0^2) are calculated for determination of $r_m{}^2$. The metric $r_m{}^2$ does not consider the differences between individual responses and the training set mean and thus avoids overestimation of the quality of prediction due to a wide response range (Y-range). Initially, the $r_m{}^2$ metric was used for the external validation using a test set, but later it was used also for the training set validation (internal validation) using LOO-predicted values. It has been shown that $r^2_{m(LOO)}$ and $r^2_{m(test)}$ might serve as stricter metrics than Q^2 and $R^2{}_{pred}$ respectively, especially for data sets with wide range of response values, as the classical metrics compare the *PRESS* values with the sum of squared deviations of individual observed values from the training set mean (Roy et al., 2009).

For the calculation of the $r_m{}^2$ metrics, we have initially arbitrarily used the observed response values in the y-axis and predicted values in the x-axis. However, the opposite may also be done. But this will result in a different value of the $r_m{}^2$ metric (i.e., $r_m{}'^2$) unless the predictions are perfect, i.e., when there is no intercept in the least squares regression line correlating observed and predicted values. This is because of the fact that the correlation between the observed (y) and predicted (x) values is same to that between the predicted (y) and observed (x) values in presence of an intercept of the corresponding least squares regression lines. However, this is not true when the intercept is set to zero. The parameters k and k' indicate the slopes of the regression lines through origin in the former and latter cases respectively (Fig. 1).

The following equations are employed for the calculation of r^2, r_0^2, $r_0'{}^2$, k and k' (Sachs, 1982)).

$$r^2 = \frac{\left[\sum(Y_{obs} - \overline{Y_{obs}})(Y_{pred} - \overline{Y_{pred}})\right]^2}{\sum\left(Y_{pred} - \overline{Y_{pred}}\right)^2 \times \sum\left(Y_{obs} - \overline{Y_{obs}}\right)^2} \tag{2}$$

$$r_0^2 = 1 - \frac{\sum\left(Y_{obs} - k \times Y_{pred}\right)^2}{\sum\left(Y_{obs} - \overline{Y_{obs}}\right)^2} \tag{3}$$

$$r_0'^2 = 1 - \frac{\sum\left(Y_{pred} - k' \times Y_{obs}\right)^2}{\sum\left(Y_{pred} - \overline{Y_{pred}}\right)^2} \tag{4}$$

$$k = \frac{\sum(Y_{obs} \times Y_{pred})}{\sum\left(Y_{pred}\right)^2} \tag{5}$$

$$k' = \frac{\sum(Y_{obs} \times Y_{pred})}{\sum\left(Y_{obs}\right)^2} \tag{6}$$

In Eqs. (2)–(6), Y_{obs} and Y_{pred} are experimental and predicted Y responses respectively. As the formulae for the calculation of r^2, r_0^2, $r_0'{}^2$, k and k' are established ones, how can one come to the conclusion that the method of RTO is not optimal for external validation just due to that Excel and SPSS have given different results for the above mentioned parameters? There may be a problem with a particular software in computation of r_0^2 and $r_0'{}^2$, but it has nothing to do with the computation of $r_m{}^2$ metrics and its use in external validation. We have never mentioned in our papers about any specific software like Excel or SPSS for calculation of established $r_m{}^2$ metrics, rather referred to the above formulae as are thoroughly discussed by Roy et al. (2012).

4. Results and discussion

The computation of $r_m{}^2$ metrics is not dependent on any specific software, rather this can be done based on the above formulae as already discussed in different papers on $r_m{}^2$ metrics (Ojha et al., 2011; Roy et al., 2012). Shayanfar and Shayanfar (2014) mentioned that "*Excel delivers two different values for r_0^2 and $r_0'^2$ (−0.06 and 0.73, respectively) while SPSS gives only one value for r_0^2 (0.95), although there is no significant difference between K and K′ values*" for their stated example dataset [y_i (experimental) = 1, 2, 3, 4, 5, 6 and y_i (calculated) = 5, 6, 7, 8, 9, 10]. However, on using the above