Available online at www.sciencedirect.com

ScienceDirect

Journal of Molecular Graphics and Modelling 26 (2008) 1315–1326

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

The importance of the domain of applicability in QSAR modeling

Shane Weaver, M. Paul Gleeson *

Computational & Structural Chemistry, GlaxoSmithKline Medicines Research Centre, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY, United Kingdom

Received 21 November 2007; received in revised form 11 January 2008; accepted 11 January 2008
Available online 18 January 2008

Abstract

The domain of applicability is an important concept in quantitative structure activity relationships (QSAR) that allows one to estimate the uncertainty in the prediction of a particular molecule based on how similar it is to the compounds used to build the model. In this paper we discuss this important concept, providing details of the development and application of a method to compute the domain of applicability within model descriptor space and structural space as defined by daylight fingerprints.

The importance of the domain of applicability is illustrated using five QSAR models generated on plasma protein binding and P450 inhibition datasets. Such methodologies will be shown to offer us a means to monitor the performance of QSARs over time, providing us both with a way to estimate the accuracy of a given prediction and identify when a model needs to be rebuilt.

Keywords: Domain of applicability; QSAR; Cytochrome P450; Plasma protein binding; PLS; Neural network; ADMET

1. Introduction

A consideration of the developability characteristics of new chemical entities (NCEs) has become increasingly important in drug discovery in the last two decades. This is driven by the fact that ~60% of drugs fail [1–3] for different ADMET reasons (absorption, distribution, metabolism, excretion and toxicity) leading to an increasing demand for in vivo, in vitro and in silico methods to screen lead compounds much earlier on in the drug discovery process.

Quantitative structure activity relationships (QSARs) [4–7] have become an important component in the compound design and progression process since they represent a much cheaper, rapid alternative to the medium throughput in vitro and low throughput in vivo assays which are generally restricted to later in the discovery cascade. A QSAR is essentially a mathematical equation that is determined from a set of molecules with known activities using computational approaches. The exact form of the relationship between structure and activity can be determined using a variety of statistical methods and computed molecular descriptors and this equation is then used to predict the activity of new molecules.

Early QSARs pioneered by Hanch and Fugita [8] consisted of relatively small number of molecules of a given chemotype being used to derive a simple linear equation to predict the next molecule in the series to be synthesised. The advantage of this approach was that the terms in the equation were generally simple and easily interpretable, while the kinds of molecules being predicted were generally very similar to those that were already synthesised, giving the user greater confidence in the model predictions. In contrast, over the past decade an increasing number of QSARs have been reported based on large, diverse datasets, commonly termed global models, which are considered more reliable at predicting diverse structures than QSARs built on small datasets of low diversity [9–13]. These models are often built using complex statistical methods, and large numbers of often sparsely populated geometrical and electrotopological descriptors [14–17], and while this may allow for a more versatile description of molecular structure and a reliable way to relate structure to activity, the multi-dimensional space defined by such a model will become increasingly complex and fragmented.

Within the pharmaceutical industry the chemotypes being synthesised at any given time depends on several factors such as the biological targets being pursued and the hits identified from

* Corresponding author. Tel.: +44 1438 768682; fax: +44 1438 763352.
E-mail address: paul.x.gleeson@gsk.com (M.P. Gleeson).

1093-3263/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2008.01.002

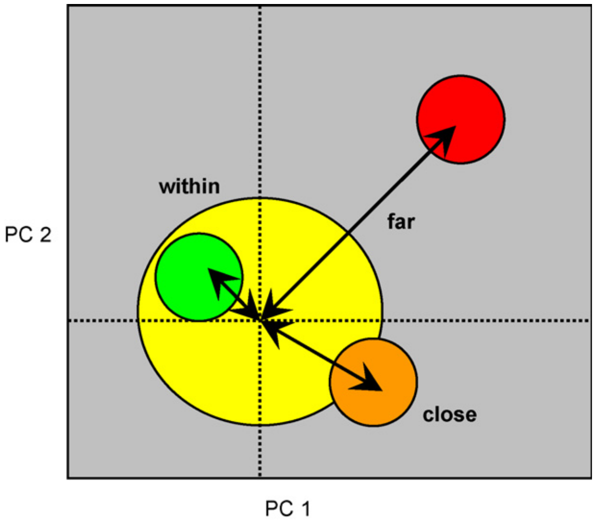

Fig. 1. A graphical illustration of the domain of applicability in principal component (PC) space. The QSAR model training set is represented by the yellow circle. Query molecules are coloured as follows: within the training space (green), close to model space (orange) and far (red). Query compounds predicted further from training model space would be expected to be less reliably predicted.

screening. This has implications for the prediction of new chemotypes not present in a QSAR model training set since these may occupy an area of model space that is not well represented (Fig. 1). Corporate collections are constantly moving further from historical chemical space meaning predictions from QSAR models developed on older, increasingly less relevant datasets will become extrapolations rather than interpolation.

Recognition of this problem in the field of QSAR can be found from the increased number of publications discussing this topic [18–26]. The methods all involve computing the similarity of the query molecules to the model training set using a variety of descriptors and distances (i.e. the so called domain of applicability), and relating this quantity to the prediction error. Readers are referred to references [19,21,23] for an introduction to the concept.

We add to the existing debate by reporting the development and application of a method to compute the domain of applicability of QSAR models, illustrating how it can be used to provide significant additional value in both local and global modeling applications. With examples derived from plasma protein binding and P450 3A4 inhibition datasets, we highlight the way in which such methods can be used to provide extra confidence in QSAR predictions.

2. Results

To illustrate the implications of the domain of applicability in QSAR modeling we have used the following methodology. Five separate QSAR models were built by splitting the respective datasets by date into three different sets as described in the experimental procedures. The earliest dated set was split into training and test sets at random, according to one of the standard practices in QSAR validation, meaning the two datasets are essentially mirror images of each other. The performance of the model on the test set represents a best-case scenario and deterioration in performance over time, and evolving chemistry might be expected. To quantify the deterioration in predictive performance we use the remaining molecules synthesised and tested over the course of at least 1 year following the completion of the model building process (validation set 1 and validation set 2).

2.1. Global plasma protein binding QSAR model

In this first example we study the effect of the domain of applicability on a QSAR model built on plasma protein binding data using a linear statistical technique called PLS regression, combined with relatively simple and interpretable 1D and 2D descriptors. Before modeling, the %bound values were transformed into the more appropriate log K scale (log(%-bound/%free)). With the exception of the newly obtained validation set 2, this model has been discussed in detail elsewhere so only a brief description is given here to facilitate a discussion of the domain of applicability calculations [24]. The QSAR, or quantitative structure property relationship (QSPR) model to be more precise, has a moderate r_0^2 of 0.56 (correlation to the line of unity—explaining 56% of the total variation), and an equivalent r^2 (regression line correlation) since this is a fitted relationship based on the 685 training set compounds with a slope of 1 and intercept of 0 (Table 1). The cross-validated q^2 is of similar magnitude at 0.54 suggesting the model is internally consistent. Note a random model prediction would have a root mean square (RMSE) $\geq$ standard deviation (S.D.).

The good model performance on the training set is no guarantee that a model will be predictive on future datasets [27]. We have therefore employed the three additional datasets discussed above (test, validation 1 and validation 2), to assess the utility of the QSPR model, each of which representing an increasingly difficult test for the model due to the increase in time, and structural diversity. Additionally, a fourth literature-derived set was available to assess the protein binding model which is also discussed.

The test set, randomly selected from the training set, is reasonably well predicted by the model. The prediction error, as given by the RMSE. Mean or median errors are comparable with those of the training set, however, the r_0^2 is considerably lower at 0.48. This is because the line of best fit though the data has a slope 0.95 and an intercept of 0.11. The Pearson's correlation coefficient (r^2) is comparable at 0.58 indicating the model has a good ranking capability. Validation set 1, consisting of data measured up to 6 months after the model was built, is less well predicted by the model again ($r_0^2 = 0.50$). The prediction error has increased while the ranking ability of the model has also decreased. Similarly validation set 2, consisting of data measured between 6 months and 1 year after the model was built, displays a further reduced r_0^2 at 0.40. The error as given by the RMSE consistent with the training and test set however, the mean and median are the largest of all the sets indicating the errors are not normally distributed making statistics requiring such normality less reliable. The final