Exploring *In Silico* Prediction of the Unbound Brain-to-Plasma Drug Concentration Ratio: Model Validation, Renewal, and Interpretation

SRINIDHI VARADHARAJAN,[1,2] SUSANNE WINIWARTER,[3] LARS CARLSSON,[3] OLA ENGKVIST,[2] AJAY ANANTHA,[2,4] THIERRY KOGEJ,[2] MARKUS FRIDÉN,[5] JONNA STÅLRING,[3] HONGMING CHEN[2]

[1]Department of Biology, Lund University, Lund SE-22100, Sweden
[2]Chemistry Innovation Centre, Discovery Sciences, AstraZeneca R&D Mölndal, Mölndal SE-43183, Sweden
[3]Computational ADME and Safety, Drug Safety and Metabolism, AstraZeneca R&D Mölndal, Mölndal SE-43183, Sweden
[4]Department of Chemistry and Molecular Biology, University of Gothenburg, Göteborg SE-40530, Sweden
[5]RIA Innovative Medicines, AstraZeneca R&D Mölndal, Mölndal SE-43183, Sweden

Received 7 October 2014; revised 14 November 2014; accepted 18 November 2014

Published online in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/jps.24301

ABSTRACT: Recently, we built an *in silico* model to predict the unbound brain-to-plasma concentration ratio ($K_{p,uu,brain}$), a measure of the distribution of a compound between the blood plasma and the brain. Here, we validate the previous model with new additional data points expanding the chemical space and use that data also to renew the model. The model building process was similar to our previous approach; however, a new set of descriptors, molecular signatures, was included to facilitate the model interpretation from a structure perspective. The best consensus model shows better predictive power than the previous model ($R^2 = 0.6$ vs. $R^2 = 0.53$, when the same 99 compounds were used as test set). The two-class classification accuracy increased from 76% using the previous model to 81%. Furthermore, the atom-summarized gradient based on molecular signature descriptors was proposed as an interesting new approach to interpret the $K_{p,uu,brain}$ machine learning model and scrutinize structure $K_{p,uu,brain}$ relationships for investigated compounds. © 2014 Wiley Periodicals, Inc. and the American Pharmacists Association J Pharm Sci

Keywords: ADME; blood–brain barrier; computational ADME; drug transport; unbound brain-to-plasma concentration ratio; machine learning; in silico modeling

INTRODUCTION

The blood–brain barrier (BBB) is a natural defense mechanism evolved to protect the delicate and sensitive brain. Consequently, it poses a major hurdle for drugs targeting the central nervous system (CNS).[1] The BBB is characterized by the presence of tight junctions that impede paracellular permeation. Additionally, the transcellular transport of more lipophilic compounds is hindered by highly expressed efflux transporters, so that the CNS is well protected from potentially harmful xenobiotics. P-glycoprotein, breast cancer-resistance protein, and multidrug-resistance protein transporters are the most vital efflux transporters at the BBB relevant for drug disposition.[2] On the contrary, influx transporters are present to ensure that compounds that are essential for the brain, for example, nutrients, can pass through the membrane. Therefore, it is the interplay of influx/efflux transporters at the BBB interface that regulates the transcellular movement of molecules across the membrane.[3]

A centrally acting drug has to cross the BBB in sufficient amount to elicit the required pharmacological effect in the CNS. For peripherally acting drugs, on the contrary, it may be advantageous to be kept out of the brain to avoid undesired side effects. Therefore, understanding the likely brain exposure for a compound in early discovery phase is crucial.

Correspondence to: Hongming Chen (Telephone: +46-31-7065285; Fax: +46-31-7763700; E-mail: hongming.chen@astrazeneca.com)

This article contains supplementary material available from the authors upon request or via the Internet at http://onlinelibrary.wiley.com/.

Journal of Pharmaceutical Sciences

Earlier, structure–brain exposure studies have largely focused on the total brain-to-plasma concentration ratio, denoted by $K_{p,brain}$ or in its logarithmic form, log BB.[4] $K_{p,brain}$[5] is described by the following equation (Eq. 1).

$$K_{p,brain} = \frac{A_{brain}}{C_p} \quad (1)$$

A_{brain} is the total amount of drug in the brain per unit tissue weight and C_p is the total concentration of the drug molecule in the blood plasma. One of the first attempts toward QSAR modeling of log BB was published by Young et al.[6] who correlated log BB with $\Delta\log P$ in a series of 20 antihistamine molecules. This was followed by several modeling attempts trying to use properties such as lipophilicity, polar surface area (PSA), hydrogen binding, and so on to predict log BB.[7–11]

However, the total amount of drug in the brain does not necessarily reflect the relevant drug concentration that is responsible for the efficacy.[5] Following the free drug hypothesis, it has been proposed that it is the unbound or free drug concentration in the brain interstitial fluid (ISF) rather than the total drug concentration in the brain, which is driving the pharmacodynamic response.[12,13] The unbound brain-to-plasma concentration ratio, $K_{p,uu,brain}$,[13–15] is a way to measure the free drug concentration in the brain in relation to the free concentration in the blood plasma. It presents information on the passive diffusion and active influx/efflux occurring at the BBB interface and is thus a relevant measure of brain uptake of drugs.[12]

$K_{p,uu,brain}$ is defined by Eq. 2.[13]

$$K_{p,uu,brain} = \frac{C_{u,brainISF}}{C_{u,p}} \quad (2)$$

$C_{u,brainISF}$ denotes the unbound concentration of drug in brain ISF and $C_{u,p}$ denotes the unbound drug concentration in the blood plasma. $C_{u,brainISF}$ can be determined directly in one *in vivo* experiment using microdialysis. However, this is a somewhat laborious method with technical limitations, for example, the method is not well suited to measure highly lipophilic compounds. Therefore, surrogate methods have been proposed to estimate the values of $K_{p,uu,brain}$ indirectly. One alternative way of determining $K_{p,uu,brain}$ involves combining the total brain-to-plasma concentration ratio $K_{p,brain}$ measured *in vivo* with the unbound volume of distribution ($V_{u,brain}$) and the unbound fraction of drug in the plasma ($f_{u,p}$) measured *in vitro* using brain slices[16] and equilibrium dialysis methods,[17] respectively (Eq. (3)).

$$K_{p,uu,brain} = \frac{K_{p,brain}}{V_{u,brain} f_{u,p}} \quad (3)$$

Note that although $K_{p,uu,brain}$ is assessed from the three directly measurable components $K_{p,brain}$, $f_{u,p}$, and $V_{u,brain}$, it is mechanistically not dependent on any of them.[15]

The first QSAR modeling study using $K_{p,uu,brain}$ data was described by Fridén et al.,[15] based on $K_{p,uu,brain}$ measurements of 41 marketed drugs. In 2011, we built a predictive model[18] by extending the Fridén dataset with a set of in-house compounds. The models were validated and shown to have decent continuous value predictions along with a good classification performance on an external test set.

Since 2011, 99 additional compounds have been measured in-house, which expands the chemical space covered in the earlier model. Because it was shown that validating and updating QSAR models regularly can improve the accuracy of a QSAR model,[19] we found it was time to revisit the model. Additionally, a new set of descriptors, the signature descriptors, was found to be a powerful tool for building QSAR models, in combination with the support vector machine (SVM) algorithm.[20] In this paper, we describe the validation of our earlier model, how the addition of the new data points improves the model and, more importantly, how the atom-based gradients of signature descriptors for SVM model can be used to identify functional groups that possibly influence $K_{p,uu,brain}$ in individual compounds.

METHODS

Experimental Data

The exact procedure to determine $K_{p,uu,brain}$ values was described earlier.[15] In short, the data are derived from three different experiments, which are combined using Eq. 3: an *in vivo* determination of the brain–blood ratio ($K_{p,brain}$) in rat and the *in vitro* assessment of binding properties in both brain ($V_{u,brain}$) and plasma ($f_{u,p}$). The *in vivo* experiment comprises a 4-h constant rate infusion in Sprague–Dawley rats with up to three drugs administered. Terminal sampling of blood and brain tissue was performed under isoflurane anesthesia. The total concentration in the brain, A_{brain}, was reduced by 0.8% of the drug plasma concentration to approximately correct for drug in the residual blood.[21] Binding to brain tissue was determined as $V_{u,brain}$ *in vitro* in brain slices[16] and plasma protein binding measured by equilibrium dialysis.[17]

Dataset

All experimental values were obtained from the corporate database. The datasets used in the present study are described as follows.

Dataset 1

Dataset 1 comprises the data used for the $K_{p,uu,brain}$ predictive model built in 2011. Dataset 1a consists of 247 compounds with $K_{p,uu,brain}$ data (values for $K_{p,brain}$, $V_{u,brain}$, and $f_{u,p}$ are all available) used for the direct model. Datasets 1b, 1c, and 1d comprise 506, 473, and 3235 compounds with measured values of $K_{p,brain}$, $V_{u,brain}$, and $f_{u,p}$, respectively. These datasets were employed for the indirect models. 73 compounds (30%) were randomly picked from dataset 1a as a test set (test set 1).[18]

Dataset 2

Since 2011, additional in-house data were accumulated for all the parameters. We collected 99 compounds for which the values of $K_{p,brain}$, $V_{u,brain}$, and $f_{u,p}$ were available. These 99 compounds comprise dataset 2, a temporal validation set for the old $K_{p,uu,brain}$ model. Further, additional data were also available for each of $K_{p,brain}$, $V_{u,brain}$, and $f_{u,p}$, with 215, 736, and 2520 new data points, respectively.

Dataset 3

Dataset 3 contains the combined datasets with all available data. Dataset 3a consists of 346 compounds with known $K_{p,uu,brain}$ data. Dataset 3b, 3c, and 3d consist of 721, 1209, and 5755 compounds with $K_{p,brain}$, $V_{u,brain}$, and $f_{u,p}$ values, respectively. Dataset 3a was randomly divided into training and test set in a 7:3 ratio and the procedure repeated 10 times. In each run, the test set consisted of 104 compounds and these 104 compounds were removed from the $K_{p,brain}$, $V_{u,brain}$, and $f_{u,p}$ sets (dataset 3b, 3c, and 3d) to obtain the respective training sets.

Descriptor Sets

AZ Descriptors

This is a set of 196 physicochemical descriptors comprising important properties such as lipophilicity, hydrogen bonding, molecular weight, polar surface area, and so on, calculated using an in-house program.[22–24]

Signature Descriptors

Signature descriptors refer to atom-based descriptors that describe the extended valence of the atoms in the molecules. In this case, a molecule is defined in terms of a set of canonical subgraphs that represent all the atoms that are at a predefined distance (often called height) from the central atom in consideration.[25] Thus, each investigated molecule is associated with a vector whose components are the occurrence of the particular signature in the structure of the molecule. The current study generated signatures of height between 0 and 3,