European Journal of Medicinal Chemistry 44 (2009) 2452–2458

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Original article

Prediction of milk/plasma concentration ratios of drugs and environmental pollutants

Michael H. Abraham*, Javier Gil-Lostes, Mohammad Fatemi

Department of Chemistry, University College London, 20 Gordon Street, London WC1H OAJ, UK

ARTICLE INFO

Article history:
Received 14 August 2008
Received in revised form
4 January 2009
Accepted 9 January 2009
Available online 20 January 2009

Keywords:
Milk
Milk/plasma ratios
Drugs
Pollutants
Linear free energy relationship
Artificial neural networks

ABSTRACT

A large database of milk/plasma ratios, *M*/*P*, for 179 drugs and hydrophobic environmental pollutants has been constructed from literature data. Application of linear analyses shows that drugs preferentially partition into the aqueous and the protein phases of milk, but that the pollutants partition into the fat phase. No useful linear equation could be obtained for the entire 179 compound data set, but an artificial neural network with only five Abraham descriptors as input resulted in errors in $\log(1 + M/P)$ of only 0.0574, 0.116 and 0.093 log units for a training set of 135 compounds, an internal test set of 22 compounds and an external test set of 22 compounds respectively. These errors correspond to 0.203, 0.193 and 0.334 log units respectively when transformed into errors in $\log(M/P)$.

1. Introduction

For many years there has been concern over the presence of drug contaminants in human milk, as set out in recent reviews [1–4], but recently there has been additional concern over environmental pollutants in human milk, see for example [5–7]. Ito et al. [2,8] have proposed an 'Exposure Index' that relates the milk to plasma ratio, *M*/*P*, the milk intake, A, and the Infant Drug Clearance to a time averaged drug exposure level,

$$\text{Exposure Index}\,(\%) = (100 \times M/P \times A)/\text{Infant Drug clearance} \quad (1)$$

A key parameter, *M*/*P*, is the ratio of drug or pollutant concentration in milk to that in plasma. Hence any method that can be used to predict *M*/*P* ratios would be of very considerable value.

Not surprisingly, there have been numerous attempts to predict these ratios. However, it is quite difficult to compare results. Some authors use *M*/*P* itself, whereas other authors use $\log(M/P)$, many authors quote only the regression coefficient, *R*, or R^2, and do not give details of the standard error or the root mean square error (much more useful statistics), and some authors erroneously refer to fits of *M*/*P* or of $\log(M/P)$ as 'predictions'. In this work we make a distinction between statistics of fitting a set of values to some equation or algorithm, and statistics that refer to the prediction of values in some external test set that has not been used to set up the equation or algorithm used to obtain the predictive values.

Even aside from the way that results are presented, the actual system is very complicated. Fleishaker et al. [9] pointed out that a drug in plasma and in milk could exist either as the free drug or as protein-bound drug, and a drug in plasma and in milk could exist as a neutral species or as an ionised species depending on the pK_a of the drug and the pH. Furthermore, a drug could partition into the aqueous phase of milk or could partition into the separate fat phase of milk. Atkinson and Begg [10] published the first comprehensive analysis of *M*/*P* ratios, based on the suggestions of Fleishaker et al. [9]. They [9] set out plasma protein binding and milk protein binding for 14 drugs, and obtained an equation that related the two. Then milk protein binding could be estimated for any drug for which plasma protein binding was known; the two protein binding values were used as part of a fitting algorithm. Atkinson and Begg [10] also corrected for ionisation of acids or bases, using the Henderson–Hasselbach equation and taking the pH of plasma as 7.4 and that of milk as 7.2 (although other workers have used 7.0 as the pH of milk [11]). There are a number of assumptions in this ionisation correction. (a) There is an equilibrium between the unionised species in milk and the unionised species in plasma, but there is no equilibrium between the ionised species in milk and the ionised species in plasma; there seems to be no evidence for this assumption at all. (b) The pK_a of a drug in plasma is the same as the

* Corresponding author. Tel.: +44 020 7679 4639; fax: +44 020 7679 7463.
E-mail address: m.h.abraham@ucl.ac.uk (M.H. Abraham).

0223-5234/$ – see front matter © 2009 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2009.01.009

pK_a in milk (taken as the pK_a in water). As far as we know, no pK_a of any drug has been determined in plasma, and so there is no basis for taking pK_a in plasma as the same as that in water. The method of Atkinson and Begg [10,12,13] was examined by Larsen et al. [14] who concluded that it had little predictive power, but the conclusions of Larsen et al. have been challenged by Doogue et al. [15] and by Ilett [16].

Other workers have avoided the protein binding and the pK_a problems altogether. Agatonovic-Kustrin et al. [17] examined a set of 60 drug compounds, and calculated 61 descriptors for each drug. This was reduced to 26 descriptors for each drug in the final artificial neural network (ANN) which was applied to values of log(M/P). The set of 60 drugs was divided into a training set and an internal validation set (50 drugs) and an independent external test set of 10 drugs to test the predictive ability of the ANN. The root mean square errors (RMSE) in log(M/P) were 0.590 for the training set, 0.900 for the internal validation set and 0.425 for the external test set; the result is rather peculiar because errors in predictions for an external test set are invariably larger than errors in fitting a training set. In a later paper [18], the same workers examined a larger set of 123 drugs and applied an ANN to M/P ratios themselves. Nine calculated descriptors were used, but no statistics at all were given. Katritzky et al. [19] investigated M/P ratios of a set of 115 drugs, using log(M/P) as the dependent variable. They started with 850 descriptors for each drug and reduced this to the best 7 descriptors that were used in a multiple linear regression. After eliminating 15 drugs, a training set of 67 drugs could be fitted with an error of 0.324 log units, and an independent test set of 33 drugs could be predicted with an error of 0.332 log units. Zhao et al. [20] used a support vector machine, SVM, method to analyze M/P ratios for 126 drugs. The only statistic they gave was an 'accuracy' of 90.48%, which refers to classification into two sets, Class 1 with $0 > M/P > 0.1$ and Class 2 with $1 > M/P > 0.1$; however this classification appears to be at odds with data in their Table 1.

Due to the wide applicability of linear free energy relationships (LFERs) in a large number of areas, it seemed useful to start our investigations on the linear modeling and prediction of M/P value using LFER methods. Although the multiple linear regression analysis (MLRA) that is used in the implementation of LFERs is a very convenient method of analysis, it is limited (as the name implies) to linear processes. When nonlinear phenomena are significant to some extent within the data investigated, LFERs are no longer the appropriate method of analysis, and nonlinear modeling techniques such as artificial neural networks (ANNs) are necessary in order to build an accurate and reliable model. ANN has recently gained much popularity in dealing with nonlinear relationships [21]. A detailed description of the theory behind a neural network has been adequately described elsewhere [22]. An ANN is a biologically inspired computer program designed to learn from data in a manner of emulating the learning pattern in the brain. Most ANN systems are very complex high dimension processing systems. The relevant principle of supervised learning in an ANN is that it takes numerical inputs (the training data) and transfers them into desired outputs. The input and output nodes may be connected to the 'external world' and to other nodes within the network. The way in which each node transforms its input depends on the so-called 'connection weights' or 'connection strength' and bias of the node, which are modifiable. The output values of each node depend on both the weight strength and bias values. For the present purpose, the great power of ANNs stems from the fact that it is possible to train them. Training is done by continually presenting the networks with known inputs and outputs and modifying the connection weights and biases between the individual nodes. This process is continued until the output nodes of the network match the desired outputs to a stated degree of accuracy. Training of the ANN can be performed by using a back-propagation algorithm. In order to train the network using a back-propagation algorithm, the differences between the ANN output and its desired value are calculated after each training iteration and the values of weights and biases modified by using these error terms. In the MLR method, the analysis is limited to a certain number of possible interactions, but in the ANN method more terms can be examined for interactions between features. ANNs are capable of recognizing nonlinear relationships between inputs and outputs. In addition, the ANN can use qualitative as well as quantitative inputs, and does not require an explicit relationship between the inputs and the outputs.

It seems therefore that there is still scope for analyses of M/P or log(M/P) values for drugs, using both LFER and ANN methods, and for investigating if the same methods can be used for environmental pollutants; to date, the latter have not been studied at all.

2. Methods

The plasma/milk system is very complicated, and it is possible that simple linear equations for M/P or for log(M/P) might not be very successful. However, as a start we used the same method that we have previously employed [23–27] for partitions from blood or plasma to various biological systems. The method uses the linear free energy relationship, LFER, shown as Eq. (2).

$$\mathrm{SP} = c + eE + sS + aA + bB + vV \quad (2)$$

In Eq (2) SP is the dependent variable, for example M/P or log(M/P), and the independent variables are properties of drugs and environmental pollutants (solutes) as follows [28,29]. E is the solute excess molar refractivity in units of $(\mathrm{cm^3\,mol^{-1}})/10$, S is the solute dipolarity/polarizability, A and B are the overall or summation hydrogen bond acidity and basicity, and V is the McGowan volume in units of $(\mathrm{cm^3\,mol^{-1}})/100$. Even if Eq. (2) itself is not successful, the five descriptors might be useful.

In addition to Eq. (2) we used an ANN that can deal with nonlinear processes, as might be the plasma–milk process. As inputs we used the five solute descriptors shown in Eq. (2) as calculated by the PharmaAlgorithms software package 'Absolv' [30]. The values of M/P that we have used [12,18–20,31–101] both for drugs and environmental pollutants are given in Table 1, together with the calculated Absolv descriptors.

3. Results and discussion

3.1. LFER methods

The plasma to milk system is very complicated, even in terms of the separate phases of milk as shown in Fig. 1. Because the volumes of the phases are not the same, the relationship between the three equilibrium constants and the overall equilibrium constant for plasma to milk, M/P, is

$$M/P = K_{pf} \times V_f/V_m + K_{pp} \times V_p/V_m + K_{pw} \times V_w/V_m \quad (3)$$

In Eq. (3), V_f/V_m, V_p/V_m and V_w/V_m are the ratios of the volumes of the fat phase, the protein phase and the aqueous phase to the total milk phase (so that $V_f + V_p + V_w = V_m$). It is rather evident, just from Eq. (2) that linear equations are not likely to yield satisfactory results. When we applied Eq. (1) to the plasma to milk system, using M/P or log(M/P) or log($1 + M/P$) we obtained no useful equations. We therefore turned to the nonlinear method of artificial neural networks.